

University of Toronto  
Department of Economics



Working Paper 375

Bootstrap-based Bandwidth Selection for Semiparametric  
Generalized Regression Estimators

By Chuan Goh

October 09, 2009

# Bootstrap-based Bandwidth Selection for Semiparametric Generalized Regression Estimators

Chuan Goh\*

30 September 2009

---

\*The author is grateful to Oliver Linton and to three referees for their constructive comments on earlier versions of this paper, and to David Jacho-Chávez, Yingying Dong and Jeff Gronek for their advice with the computations performed for the current version of this paper. The author would particularly like to acknowledge the hospitality and support of the faculty and staff of the Department of Economics at Indiana University, where much of the work on this paper was completed. This research was supported in part by the Connaught Fund of the University of Toronto. Address correspondence to: Department of Economics, University of Toronto, Max Gluskin House, 150 St. George St., Toronto, ON, Canada, M5S 3G7. [goh@economics.utoronto.ca](mailto:goh@economics.utoronto.ca), <http://www.chuangoh.org>.

## Abstract

This paper considers the problem of implementing semiparametric extremum estimators of a generalized regression model with an unknown link function. The class of estimator under consideration includes as special cases the semiparametric least-squares estimator of Ichimura (1993) as well as the semiparametric quasi-likelihood estimator of Klein and Spady (1993). In general, it is assumed to involve the computation of a nonparametric kernel estimate of the link function that appears in place of the true, but unknown, link function in the appropriate location in a smooth criterion function. The specific question considered in this paper concerns the practical selection of the degree of smoothing to be used in computing the nonparametric regression estimate. This paper proposes a method for selecting the smoothing parameter via resampling. The particular method suggested here involves using a resample of smaller size than the original sample. Specific guidance on selecting the resample size is given, and simulation evidence is presented to illustrate the utility of this method for samples of moderate size.

*JEL Classification: C14*

**KEYWORDS:** Bandwidth selection, semiparametric, single-index model, bootstrap,  $m$ -out-of- $n$  bootstrap, kernel smoothing

# 1 Introduction

This paper is concerned with the general issue of smoothing parameter selection for nonparametric estimators that are used as components of a semiparametric estimator. In particular, the focus is on the implementation of members of a class of semiparametric  $M$ -estimator involving both the estimation of an infinite-dimensional nuisance parameter by some method of smoothing and the computation of an estimate of a finite-dimensional parameter of interest by minimization of a smooth objective function. It is assumed that the nonparametric estimator of the infinite-dimensional nuisance parameter appears within the objective function in place of the true, but unknown, infinite-dimensional nuisance parameter. Examples of the many semiparametric estimators that can be included in this framework include the proposals of Robinson (1988); Powell et al. (1989); Ichimura (1993) and Klein and Spady (1993).

The specific class of estimator considered in this paper is concerned with the estimation of a conditional mean possessing a single-index structure in the conditioning variables that enter the regression via an unknown link function. An obvious, though perhaps uninteresting, example of a single-index model belonging to this class is the linear regression model. Other cases include models involving dependent variables characterized by censoring and sample selectivity. In general, the parameter of interest in these cases is naturally the vector of index coefficients, while the link function is treated as an infinite-dimensional nuisance parameter. Ignorance of the link function is assumed in this paper to be accommodated using a nonparametric kernel regression method of the Nadaraya-Watson type. The vector of index coefficients is estimated by means of an explicit optimization procedure, as in Ichimura (1993) or Klein and Spady (1993). The first-order asymptotic distribution theory for semiparametric estimators of this class has been described by Andrews (1994), Newey (1994) and Sherman (1994).

The theory developed in this paper is largely concerned with the higher-order asymptotic properties of the particular class of semiparametric  $M$ -estimator under consideration. The reason for this is that the asymptotic distribution of the normalized estimator does not actually depend on the bandwidth used to construct the kernel regression estimates of the link function. Asymptotic approaches to bandwidth selection in this setting will therefore depend on the use of higher-order distributional approximations.

This paper considers the specific context of semiparametric  $M$ -estimation of a generalized regression model described above and proposes a new method of selecting the bandwidth in applications. The approach taken in this paper

was inspired by a suggestion of Horowitz (2009, §2.7) and involves the use of resampling—with replacement—fewer observations than are present in the original sample—the so-called “ $m$ -out-of- $n$ ” or “ $m$ -bootstrap”.<sup>1</sup>

The approach taken in this paper also complements existing methods based on resampling as many observations as exist in the original sample coupled with an explicit method of bias correction. The “manual” bias correction called for in this case arises out of the inability of the full-sample bootstrap to generate adequate approximations to the bias of the semiparametric estimator.<sup>2</sup> The approach taken in this paper avoids any need to engage in the sort of case-specific explicit bias correction required by approaches involving the full-sample bootstrap.

The remainder of the paper proceeds as follows. The following section presents the specific estimation problem considered in this paper and describes the class of semiparametric  $M$ -estimator that is being assumed to handle it. The main results of this paper appear in Section 3. In particular, higher-order asymptotic expansions of the estimator are derived under both the original and bootstrap sampling schemes and are used to motivate the method of bandwidth selection proposed here. Section 4 presents the results of simulation experiments evaluating the small-sample performance of the method of bandwidth selection described in Section 3. Section 5 concludes. Detailed proofs of the analytical results presented in Section 3 appear in the appendix.

## 2 Setup

Suppose  $\{\mathbf{Z}_i \equiv (\mathbf{X}_i^\top, Y_i)^\top : i = 1, \dots, n\}$  is a sample from a population of random  $(d + 2)$ -variates  $\mathbf{Z} \equiv (\mathbf{X}^\top, Y)^\top$ , where  $\mathbf{X}$  is a random  $(d + 1)$ -variate, and where

$$E[Y | \mathbf{X}] = G_0(\mathbf{X}^\top \boldsymbol{\beta}(\boldsymbol{\theta}_0)) \quad (1)$$

for some unknown link function  $G_0 : \mathbb{R} \rightarrow \mathbb{R}$ . As such, the data are drawn from a single-index model, which naturally subsumes the class of multiple linear regression models, Tobit models and various econometric models of discrete choice as

---

<sup>1</sup>An analogous procedure involving resampling *without* replacement (i.e., “subsampling”) should also be feasible in the context considered in this paper. In particular, resampling without replacement is the actual mode of resampling suggested in Horowitz (2009, §2.7).

<sup>2</sup>See e.g., the approach taken by Nishiyama and Robinson (2005) for the case of semiparametric estimators of density-weighted average derivatives.

special cases.<sup>3</sup>

In this paper it is explicitly assumed that the link function in (1) is unknown and that one takes a semiparametric approach to estimating the parametric index. In addition, it is assumed that the  $(d + 1)$ -dimensional vector of index coefficients in (1) has the form

$$\boldsymbol{\beta}(\boldsymbol{\theta}_0) \equiv (1, \boldsymbol{\theta}_0^\top)^\top$$

for identifiability.

Let  $K$  denote a univariate kernel function for density estimation and suppose that  $\{h_n\}$  is a sequence of non-negative real numbers with  $h_n \rightarrow 0$  as  $n \rightarrow \infty$ . Define  $K_{h_n} \equiv K\left(\frac{t}{h_n}\right)$  and set

$$\hat{g}_n^{(i)}(u|\boldsymbol{\theta}) \equiv \frac{\sum_{j \neq i} Y_j K_{h_n}(u - \mathbf{X}_j^\top \boldsymbol{\beta}(\boldsymbol{\theta}))}{\sum_{j \neq i} K_{h_n}(u - \mathbf{X}_j^\top \boldsymbol{\beta}(\boldsymbol{\theta}))}; \quad (2)$$

$$\hat{f}_n^{(i)}(u|\boldsymbol{\theta}) \equiv \frac{1}{(n-1)h_n} \sum_{j \neq i} K_{h_n}(u - \mathbf{X}_j^\top \boldsymbol{\beta}(\boldsymbol{\theta})). \quad (3)$$

It follows that  $\hat{g}_n^{(i)}(u|\boldsymbol{\theta})$  is a Nadaraya-Watson estimator estimating the conditional mean

$$g_0(u|\boldsymbol{\theta}) \equiv E[Y | \mathbf{X}^\top \boldsymbol{\beta}(\boldsymbol{\theta}) = u],$$

while  $\hat{f}_n^{(i)}(u|\boldsymbol{\theta})$  estimates the marginal density  $f_0(u|\boldsymbol{\theta})$  of  $\mathbf{X}^\top \boldsymbol{\beta}(\boldsymbol{\theta})$  evaluated at the point  $\mathbf{X}^\top \boldsymbol{\beta}(\boldsymbol{\theta}) = u$ .

It is noted that there exist many possibilities for the semiparametric estimation of the finite-dimensional parameter  $\boldsymbol{\theta}_0$  in the context of a generalized regression model taking the form given in (1) above.<sup>4</sup> This paper, however, is concerned with estimates of  $\boldsymbol{\theta}_0$  having the form

$$\hat{\boldsymbol{\theta}}_n = \arg \min_{\boldsymbol{\theta} \in \mathbb{R}^d} \hat{S}_n(\boldsymbol{\theta}), \quad (4)$$

where

$$\hat{S}_n(\boldsymbol{\theta}) \equiv \frac{1}{n} \sum_{i=1}^n \rho(Y_i, \hat{g}_n^{(i)}(\mathbf{X}_i^\top \boldsymbol{\beta}(\boldsymbol{\theta})|\boldsymbol{\theta})) \hat{\tau}_{ni}(\boldsymbol{\theta}). \quad (5)$$

---

<sup>3</sup>The estimation of the parametric index and infinite-dimensional link function has been the subject of a fairly lengthy literature in econometrics and statistics more generally; a review is available in Carroll et al. (1997).

<sup>4</sup>In particular, the method of average derivatives proposed by Härdle and Stoker (1989), the sliced inverse regression method of Li (1991) and the semiparametric least squares procedure of Ichimura (1993) are amongst the most popular.

It follows that  $\hat{\boldsymbol{\theta}}_n$  is a semiparametric  $M$ -estimator, and subsumes a number of estimators previously considered in the econometric literature as special cases.<sup>5</sup>

In the context of (5), the function  $\rho : \mathbb{R}^2 \rightarrow \mathbb{R}$  denotes a known criterion function possessing a sufficient order of differentiability in the second argument, while  $\hat{\tau}_{ni}(\boldsymbol{\theta})$  denotes a data-dependent trimming function used to guarantee the  $\sqrt{n}$ -consistency of  $\hat{\boldsymbol{\theta}}_n$  in the face of observations involving values of  $f_0(\mathbf{X}_i^\top \boldsymbol{\beta}(\boldsymbol{\theta}) | \boldsymbol{\theta})$  close to zero.<sup>6</sup> In particular, we set

$$\hat{\tau}_{ni} \equiv \tau_n \left( \hat{f}_n^{(i)}(\mathbf{X}_i^\top \boldsymbol{\beta}(\boldsymbol{\theta}) | \boldsymbol{\theta}) \right)$$

for some smooth function  $\tau_n(\cdot)$  bounded on  $\mathbb{R}$  with

$$\tau_n(u) = \begin{cases} 1 & , \quad u \geq 2n^{-\zeta} \\ 0 & , \quad u \leq n^{-\zeta} \end{cases}$$

for a suitable positive constant  $\zeta$ .<sup>7</sup> The function  $\tau_n(\cdot)$  is taken to be monotonically increasing and smooth on  $\mathbb{R}$  and taking values in  $(0, 1)$  whenever  $n^{-\zeta} < u < 2n^{-\zeta}$ .<sup>8</sup>

### 3 Main Results

The asymptotic behaviour of the class of estimator considered in this paper is derived under the following regularity conditions. In this connection, define

$$\begin{aligned} \tilde{\rho}(\mathbf{Z}_i, \boldsymbol{\theta}) &\equiv \rho(Y_i, g_0(\mathbf{X}_i^\top \boldsymbol{\beta}(\boldsymbol{\theta}) | \boldsymbol{\theta})) ; \\ \rho^{(m)}(s, t) &\equiv \frac{\partial^m}{\partial t^m} \rho(s, t), \quad m = 0, 1, 2, \dots \end{aligned}$$

Also let  $\nabla_m$  denote the  $m$ th-order partial derivative operator with respect to  $\boldsymbol{\theta}$ .

---

<sup>5</sup>In addition to the procedure of Ichimura (1993), the estimators of Robinson (1988), Powell et al. (1989) and Klein and Spady (1993) can be categorized as belonging to this class.

<sup>6</sup>Technical discussions of various trimming procedures can be found in among others, Härdle and Stoker (1989), Härdle et al. (1993), Ichimura (1993) and Sherman (1994).

<sup>7</sup>The precise restrictions on  $\zeta$  are given in Assumption 6 below.

<sup>8</sup>An example would be

$$\tau_n(u) = \frac{\exp \left[ (2n^{-\zeta} - u)^{-1} \right]}{\exp \left[ (2n^{-\zeta} - u)^{-1} \right] + \exp \left[ (u - n^{-\zeta})^{-1} \right]}.$$

**Assumption 1** (Identifiability and rate of convergence). 1.  $\boldsymbol{\theta}_0$  is an interior point of  $\Theta$ , which is in turn a compact subset of  $\mathbb{R}^d$ .

2.  $\boldsymbol{\theta}_0$  is the unique minimizer of  $E [\rho(Y, g_0(\mathbf{X}^\top \boldsymbol{\beta}(\boldsymbol{\theta}) | \boldsymbol{\theta}))]$  over  $\Theta$ .

3.  $\hat{\boldsymbol{\theta}}_n \in \Theta_n$ , where

$$\Theta_n \equiv \left\{ \boldsymbol{\theta} : \|\boldsymbol{\theta} - \boldsymbol{\theta}_0\| \leq C_0 n^{-\frac{1}{2}} \right\}$$

for some  $C_0 > 0$ .

**Assumption 2** (Data-generating process). 1. The data  $\{(\mathbf{X}_i^\top, Y_i)^\top\}$  constitute a random sample from a population  $(\mathbf{X}^\top, Y)^\top$  satisfying the single-index structure given above in (1).

2. The density function  $f_0(u | \boldsymbol{\theta})$  of  $\mathbf{X}^\top \boldsymbol{\beta}(\boldsymbol{\theta})$  and its derivatives up to fourth order are bounded on  $\mathbb{R}$  for every  $\boldsymbol{\theta} \in \Theta_n$  and for every  $\boldsymbol{\theta} \in \Theta_n$  and for every  $u \in \Lambda_n$ , where  $\Theta_n$  is as given above in the statement of Assumption 1 and

$$\Lambda_n \equiv \{u : |u| < n^c, f_0(u | \boldsymbol{\theta}) > n^{-2\zeta}, \boldsymbol{\theta} \in \Theta_n\}.$$

3.  $\sup_{u \in \mathbb{R}, \boldsymbol{\theta} \in \Theta_n} |f_0(u | \boldsymbol{\theta}) - f_0(u | \boldsymbol{\theta}_0)| \leq c \|\boldsymbol{\theta} - \boldsymbol{\theta}_0\|$

4.  $E [\|\mathbf{X}\|^6] < \infty$ .

5.  $E [|Y|^3] < \infty$  for some constant  $c > 0$ .

**Assumption 3** (Link function).

$$\sup_{u \in \Lambda_n, \boldsymbol{\theta} \in \Theta} \left| g_0^{(m)}(u | \boldsymbol{\theta}) \right| < \infty$$

for  $m = 0, 1, 2, 3, 4$ , where  $\Lambda_n$  is as given above in the statement of Assumption 2.

**Assumption 4** (Criterion function). 1.

$$E [\rho^{(1)}(Y, g_0(\mathbf{X}^\top \boldsymbol{\beta}(\boldsymbol{\theta}) | \boldsymbol{\theta}))] = 0.$$

2. For each  $\mathbf{z}$  in the support of  $\mathbf{Z} \equiv (\mathbf{X}^\top, Y)^\top$ , all mixed partial derivatives of  $\tilde{\rho}(\mathbf{z}, \cdot)$  exist to third order on  $\Theta$ .



3.

$$E \left[ \sup_{\Theta} \|\nabla_3 \tilde{\rho}(\mathbf{Z}, \boldsymbol{\theta})\| \right] < \infty.$$

4.

$$E [\|\nabla_1 \tilde{\rho}(\mathbf{Z}, \boldsymbol{\theta}_0)\|^2] < \infty$$

and

$$E [\|\nabla_2 \tilde{\rho}(\mathbf{Z}, \boldsymbol{\theta}_0)\|] < \infty.$$

5. The matrix  $E [\nabla_2 \tilde{\rho}(\mathbf{Z}, \boldsymbol{\theta}_0)]$  is positive definite.

6. For  $m \in \{0, 1, 2, 3\}$ ,

$$E \left[ \sup_{\boldsymbol{\theta} \in \Theta} [\rho^{(m)}(Y, g_0(\mathbf{X}^\top \boldsymbol{\beta}(\boldsymbol{\theta}) | \boldsymbol{\theta}))]^2 \right] < \infty.$$

**Assumption 5** (Kernel function).  $K(s)$  is a symmetric density function possessing finite moments of all orders.

**Assumption 6** (Bandwidth and trimming function). 1.  $h_n \in [\delta n^{-\frac{1}{3}}, \delta^{-1} n^{-\frac{1}{3}}]$  for some small  $\delta \in (0, 1)$ .

2.  $\tau_n(u)$  is bounded and boundedly differentiable on  $\mathbb{R}$  with

$$\tau_n(u) = \begin{cases} 1 & , \quad u > 2n^{-\zeta}; \\ 0 & , \quad u \leq n^{-\zeta}. \end{cases}$$

3.  $\zeta \in (0, \frac{2}{15})$ .

It is noted that part 1 of Assumption 1 is standard. Part 2 of Assumption 1 assumes identification, while part 3 assumes the existence of a root- $n$ -consistent estimator of  $\boldsymbol{\theta}_0$ . The conditions of Assumption 2 allow for the design vector  $\mathbf{X}$  to have discrete components as long as  $\mathbf{X}^\top \boldsymbol{\beta}(\boldsymbol{\theta})$  is continuous in a root- $n$ -neighbourhood of  $\boldsymbol{\theta}_0$ . The moment requirement on  $\mathbf{X}$  is the same as that imposed by Xia et al. (2009), and is in fact weaker than the bounded support condition for  $\mathbf{X}$  imposed in Härdle et al. (1993). Part 5 of Assumption 2 is designed to ensure the applicability of Xia et al. (2009, Lemma 6.1), which appears in the

Appendix as Lemma 1. The differentiability condition on the link function in Assumption 3 helps ensure that the bias of  $\hat{g}_n^{(i)}(u|\boldsymbol{\theta})$  as an estimator of  $g_0(u|\boldsymbol{\theta})$  is of order  $O(h_n^2)$  uniformly for  $u \in \Lambda_n$  and  $\boldsymbol{\theta} \in \Theta$ . The condition of Assumption 3 may be relaxed to a requirement of a uniformly bounded second derivative at the expense of additional analytical complications in the proofs and a smaller bandwidth. Part 1 of Assumption 4 ensures that  $\hat{\boldsymbol{\theta}}_n$  has the same asymptotic distribution as the minimizer of an idealized criterion function where  $g_0(\mathbf{X}_i^\top \boldsymbol{\beta}(\boldsymbol{\theta})|\boldsymbol{\theta})$  appears in place of its estimate  $\hat{g}_n^{(i)}(\mathbf{X}_i^\top \boldsymbol{\beta}(\boldsymbol{\theta})|\boldsymbol{\theta})$ .<sup>9</sup> Parts 2–5 of Assumption 4 are standard regularity conditions that ensure the existence of a Taylor expansion of  $\tilde{\rho}(\mathbf{z}, \boldsymbol{\theta})$  about  $\tilde{\rho}(\mathbf{z}, \boldsymbol{\theta}_0)$ . Part 6 of Assumption 6 is imposed primarily for convenience in establishing the first-order asymptotic representation of  $\sqrt{n}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0)$  given below in Theorem 1.<sup>10</sup> Assumption 5 allows for the use of the most popular smoothing kernels in empirical practice, including in particular the Gaussian kernel. Part 1 of Assumption 6 presets the rate of decay of the bandwidth to that which minimizes the order of the two largest higher-order bias terms in the stochastic expansion of  $\sqrt{n}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0)$  given below in Theorem 2. Finally, part 3 of Assumption 6 ensures that the effect of trimming is asymptotically negligible in the stochastic expansion for  $\sqrt{n}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0)$ .

For completeness, the asymptotic behaviour of the estimator to first order is explicitly stated.

**Theorem 1** (First-Order Asymptotics). *Under the conditions of Assumptions 1–6,  $\hat{\boldsymbol{\theta}}_n$  is asymptotically normal and satisfies*

$$\sqrt{n}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0) = \frac{1}{\sqrt{n}} \sum_{i=1}^n \boldsymbol{\Psi}(\mathbf{Z}_i, \boldsymbol{\theta}_0, \hat{g}_n^{(i)}) + o(1),$$

almost surely, where for each  $i \in \{1, \dots, n\}$ ,

$$\begin{aligned} & \boldsymbol{\Psi}(\mathbf{Z}_i, \boldsymbol{\theta}_0, \hat{g}_n^{(i)}) \\ & \equiv \left[ \frac{1}{n} \sum_{j=1}^n \nabla_2 \rho(Y_j, \hat{g}_n^{(i)}(\mathbf{X}_j^\top \boldsymbol{\beta}(\boldsymbol{\theta}_0)|\boldsymbol{\theta}_0)) \hat{\tau}_{nj}(\boldsymbol{\theta}_0) \right]^{-1} \\ & \quad \cdot \nabla_1 \rho(Y_i, \hat{g}_n^{(i)}(\mathbf{X}_i^\top \boldsymbol{\beta}(\boldsymbol{\theta}_0)|\boldsymbol{\theta}_0)) \hat{\tau}_{ni}(\boldsymbol{\theta}_0). \end{aligned}$$

<sup>9</sup>See e.g., Sherman (1994).

<sup>10</sup>It is verifiable for the case of the semiparametric least squares procedure of Ichimura (1993) and also for the pseudo-maximum likelihood estimator of the binary-choice model considered by Klein and Spady (1993) under the regularity conditions imposed by those authors.

*Proof.* Appendix A.2. □

An expansion of  $\Psi \left( \mathbf{Z}_i, \boldsymbol{\theta}_0, \hat{g}_n^{(i)} \right)$  about  $g_0 \left( \mathbf{X}_i^\top \boldsymbol{\beta}(\boldsymbol{\theta}_0) \mid \boldsymbol{\theta}_0 \right)$  gives immediate access to the higher-order effects of the nonparametric estimation of the link function on the asymptotic behaviour of  $\hat{\boldsymbol{\theta}}_n$ , as is summarized in the following result. The precise form of the asymptotic variance of the estimator is an immediate corollary of the higher-order expansion, which incidentally does not involve a curse of dimensionality.

**Theorem 2** (Higher-Order Asymptotics). *Recall the definition of  $\Lambda_n$  given above in the statement of Assumption 2. Let*

$$\begin{aligned} \epsilon_i(\boldsymbol{\theta}) &\equiv Y_i - g_0 \left( \mathbf{X}_i^\top \boldsymbol{\beta}(\boldsymbol{\theta}) \mid \boldsymbol{\theta} \right), \\ \sigma^2(\boldsymbol{\theta}_0) &\equiv E \left[ \epsilon_1^2(\boldsymbol{\theta}_0) \right] \end{aligned}$$

and

$$\begin{aligned} \psi(\mathbf{Z}_i, \boldsymbol{\theta}, g) &\equiv \left[ \frac{1}{n} \sum_{\{j: \mathbf{X}_j^\top \boldsymbol{\beta}(\boldsymbol{\theta}_0) \in \Lambda_n\}} \nabla_{2\rho} \left( Y_j, g \left( \mathbf{X}_j^\top \boldsymbol{\beta}(\boldsymbol{\theta}) \mid \boldsymbol{\theta} \right) \right) \right]^{-1} \\ &\quad \cdot \nabla_{1\rho} \left( Y_i, g \left( \mathbf{X}_i^\top \boldsymbol{\beta}(\boldsymbol{\theta}) \mid \boldsymbol{\theta} \right) \right). \end{aligned}$$

Denote the first and second-order derivatives of  $\psi$  with respect to  $g$  as  $\psi_g$  and  $\psi_{gg}$ , respectively, and define

$$\begin{aligned} \bar{\mu}_{gi}(\boldsymbol{\theta}_0, g_0) &\equiv E \left[ \psi_g(\mathbf{Z}_i, \boldsymbol{\theta}_0, g_0) \mid \mathbf{X}_i^\top \boldsymbol{\beta}(\boldsymbol{\theta}_0) \in \Lambda_n \right], \\ \bar{\mu}_{ggi}(\boldsymbol{\theta}_0, g_0) &\equiv E \left[ \psi_{gg}(\mathbf{Z}_i, \boldsymbol{\theta}_0, g_0) \mid \mathbf{X}_i^\top \boldsymbol{\beta}(\boldsymbol{\theta}_0) \in \Lambda_n \right]. \end{aligned}$$

Furthermore, set

$$\begin{aligned} b_0(\mathbf{Z}_1 \mid \boldsymbol{\theta}_0) &\equiv \frac{1}{2} \left[ g_0''(\mathbf{x}^\top \boldsymbol{\beta}(\boldsymbol{\theta}) \mid \boldsymbol{\theta}) + \frac{2g_0'(\mathbf{x}^\top \boldsymbol{\beta}(\boldsymbol{\theta}) \mid \boldsymbol{\theta}) f_0'(\mathbf{x}^\top \boldsymbol{\beta}(\boldsymbol{\theta}) \mid \boldsymbol{\theta})}{f_0(\mathbf{x}^\top \boldsymbol{\beta}(\boldsymbol{\theta}) \mid \boldsymbol{\theta})} \right. \\ &\quad \left. + \frac{g_0(\mathbf{x}^\top \boldsymbol{\beta}(\boldsymbol{\theta}) \mid \boldsymbol{\theta}) f_0''(\mathbf{x}^\top \boldsymbol{\beta}(\boldsymbol{\theta}) \mid \boldsymbol{\theta})}{f_0(\mathbf{x}^\top \boldsymbol{\beta}(\boldsymbol{\theta}) \mid \boldsymbol{\theta})} \right] \cdot \int s^2 K(s) ds. \end{aligned}$$

Then under the conditions of Assumptions 1–6, we have

$$\sqrt{n} \left( \hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0 \right) = \mathbf{A}_{n0} + h_n^2 \sqrt{n} \boldsymbol{\gamma}_1 + \frac{1}{h_n \sqrt{n}} \boldsymbol{\gamma}_2 + o \left( n^{-\frac{1}{5}} \sqrt{\log n} \right)$$

almost surely, where

$$\begin{aligned}\mathbf{A}_{n0} &\equiv \frac{1}{\sqrt{n}} \sum_{i=1}^n (\psi(\mathbf{Z}_i, \boldsymbol{\theta}_0, g_0) + \bar{\boldsymbol{\mu}}_{gi}(\boldsymbol{\theta}_0, g_0) \epsilon_i(\boldsymbol{\theta}_0)), \\ \gamma_1 &\equiv E [\bar{\boldsymbol{\mu}}_{g1}(\boldsymbol{\theta}_0, g_0) b_0(\mathbf{Z}_1 | \boldsymbol{\theta}_0)]\end{aligned}$$

and

$$\gamma_2 \equiv \sigma^2(\boldsymbol{\theta}_0) \int K^2(s) ds \cdot E \left[ \bar{\boldsymbol{\mu}}_{gg1}(\boldsymbol{\theta}_0, g_0) \cdot \frac{1}{f_0(\mathbf{X}_1^\top \boldsymbol{\beta}(\boldsymbol{\theta}_0) | \boldsymbol{\theta}_0)} \right].$$

*Proof.* Appendix A.3. □

**Corollary 1** (Asymptotic Variance).

$$\sqrt{n} (\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0) \xrightarrow{d} N(\mathbf{0}, \boldsymbol{\Sigma}_0),$$

where

$$\boldsymbol{\Sigma}_0 \equiv E \left[ (\psi(\mathbf{Z}_1, \boldsymbol{\theta}_0, g_0) + \bar{\boldsymbol{\mu}}_{g1}(\boldsymbol{\theta}_0, g_0) \epsilon_1(\boldsymbol{\theta}_0)) (\psi(\mathbf{Z}_1, \boldsymbol{\theta}_0, g_0) + \bar{\boldsymbol{\mu}}_{g1}(\boldsymbol{\theta}_0, g_0) \epsilon_1(\boldsymbol{\theta}_0))^\top \right].$$

It is readily seen from the conclusion of Theorem 2 that the order of the stochastic expansion of  $\sqrt{n} (\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0)$  may be minimized by equating the orders of the two higher-order bias terms; in particular, by setting  $h_n \propto n^{-\frac{1}{3}}$ .<sup>11</sup> We can moreover consider picking the bandwidth  $h_n$  to minimize the expectation of the weighted  $L^2$ -norm  $n (\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0)^\top \boldsymbol{\Sigma}_0^- (\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0)$ , where  $\boldsymbol{\Sigma}_0^-$  denotes the generalized inverse.<sup>12</sup> In particular,

$$n (\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0)^\top \boldsymbol{\Sigma}_0^- (\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0)$$

---

<sup>11</sup>It should also be possible to follow e.g., the analysis of Ichimura and Linton (2005) of the higher-order properties of the average treatment effect estimator of Hirano et al. (2003) in developing an analytical method of bias correction for the specific class of estimator dealt with in this paper. In particular, it should be possible to develop analytically a method of eliminating the term of order  $\frac{1}{h_n \sqrt{n}}$  in the stochastic expansion of  $\sqrt{n} (\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0)$ . This would allow for the use of smaller bandwidths and better estimator performance in terms of mean squared error. In this case, an analogous stochastic expansion for the bias-corrected estimator in the bootstrap domain, paralleling Theorem 3 below, would also be feasible, which would lead in turn to a bootstrap-based method of bandwidth selection for the bias-corrected estimator paralleling the method for the non-bias corrected estimator developed below.

<sup>12</sup>Analogous bandwidth selection procedures can also be developed for any scalar criterion derived from the asymptotic mean squared error of  $\hat{\boldsymbol{\theta}}_n$ .

$$\begin{aligned}
&= \left( \mathbf{A}_{n0} + \sqrt{n}h_n^2\gamma_1 + \frac{1}{h_n\sqrt{n}}\gamma_2 + o_{a.s.}\left(n^{-\frac{1}{5}}\sqrt{\log n}\right) \right)^\top \Sigma_0^- \\
&\quad \cdot \left( \mathbf{A}_{n0} + \sqrt{n}h_n^2\gamma_1 + \frac{1}{h_n\sqrt{n}}\gamma_2 + o_{a.s.}\left(n^{-\frac{1}{5}}\sqrt{\log n}\right) \right) \\
&= \left( \mathbf{A}_{n0} + \sqrt{n}h_n^2 + \frac{1}{h_n\sqrt{n}}\gamma_2 \right)^\top \Sigma_0^- \left( \mathbf{A}_{n0} + \sqrt{n}h_n^2 + \frac{1}{h_n\sqrt{n}}\gamma_2 \right) \\
&\quad + o_{a.s.}\left(n^{-\frac{1}{5}}\sqrt{\log n}\right) \\
&\equiv \mathbf{W}_n^\top \Sigma_0^- \mathbf{W}_n + o_{a.s.}\left(n^{-\frac{1}{5}}\sqrt{\log n}\right),
\end{aligned}$$

from which it follows that the optimal bandwidth should minimize  $E[\mathbf{W}_n^\top \Sigma_0^- \mathbf{W}]$ . In this connection, note that

$$\begin{aligned}
&E[\mathbf{W}_n^\top \Sigma_0^- \mathbf{W}_n] \\
&= E[\mathbf{A}_{n0}^\top \Sigma_0^- \mathbf{A}_{n0}] \\
&\quad + \left( \sqrt{n}h_n^2\gamma_1 + \frac{1}{h_n\sqrt{n}}\gamma_2 \right)^\top \Sigma_0^- \left( \sqrt{n}h_n^2\gamma_1 + \frac{1}{h_n\sqrt{n}}\gamma_2 \right). \quad (6)
\end{aligned}$$

Since the leading term  $E[\mathbf{A}_{n0}^\top \Sigma_0^- \mathbf{A}_{n0}]$  does not depend on  $h_n$ , it suffices to pick  $h_n$  to minimize the second term in (6). It follows that the optimal bandwidth is given by

$$h_{n,opt} = \left( \frac{\sqrt{(\gamma_1^\top \Sigma_0^- \gamma_2)^2 + 8(\gamma_1^\top \Sigma_0^- \gamma_1)(\gamma_2^\top \Sigma_0^- \gamma_2)} - \gamma_1^\top \Sigma_0^- \gamma_2}{4\gamma_1^\top \Sigma_0^- \gamma_1} \right)^{\frac{1}{3}} n^{-\frac{1}{3}}. \quad (7)$$

Although it is possible to derive a plug-in procedure to estimate the optimal bandwidth given in (7), the practical implementation of such a method is likely to be complicated by the need to estimate the first and second-order derivatives embedded in the expressions for  $\gamma_1$  and  $\gamma_2$ . In particular, this would require the further complication of selecting pilot bandwidths and affords no guarantee regarding the stability of the resulting estimates of  $h_{n,opt}$  produced by such a plug-in procedure. A possible alternative would be to replace the nonparametric derivative estimates with parametric estimates, leading to a rule-of-thumb bandwidth qualitatively similar to the classic method of Silverman (1986) in the context of density estimation. The drawback of such a procedure would be the strong assumptions regarding the data-generating process required for its implementation,

which seem to go against the whole rationale for resorting to a semiparametric procedure in the first place.

This paper accordingly develops a resampling-based method to estimate the optimal bandwidth  $h_{n,opt}$ . In this connection, let

$$\mathcal{X}_n \equiv \{ \mathbf{Z}_i \equiv (\mathbf{X}_i^\top, Y_i)^\top : i = 1, \dots, n \}$$

denote the original sample, and let

$$\mathcal{X}_m^* \equiv \{ \mathbf{Z}_{i,m}^* \equiv (\mathbf{X}_{i,m}^{*\top}, Y_{i,m}^*)^\top : i = 1, \dots, m \}$$

denote a sample of size  $m \leq n$  drawn randomly with replacement from  $\mathcal{X}_n$ . Define bootstrap counterparts to quantities previously considered, to wit:

$$\begin{aligned} \hat{g}_m^{*(i)}(u|\boldsymbol{\theta}) &\equiv \frac{\sum_{j \neq i} Y_{j,m}^* K_{h_m}(u - \mathbf{X}_{j,m}^{*\top} \boldsymbol{\beta}(\boldsymbol{\theta}))}{\sum_{j \neq i} K_{h_m}(u - \mathbf{X}_{j,m}^{*\top} \boldsymbol{\beta}(\boldsymbol{\theta}))}, \\ \hat{f}_m^{*(i)}(u|\boldsymbol{\theta}) &\equiv \frac{1}{(m-1)h_m} \sum_{j \neq i} K_{h_m}(u - \mathbf{X}_{j,m}^{*\top} \boldsymbol{\beta}(\boldsymbol{\theta})), \end{aligned}$$

where for  $h_m \rightarrow 0$  as  $m, n \rightarrow \infty$ ,

$$K_{h_m}(t) \equiv K\left(\frac{t}{h_m}\right),$$

where  $K$  is a smoothing kernel satisfying the conditions of Assumption 5. Also define

$$\hat{\boldsymbol{\theta}}_m^* \equiv \arg \min_{\boldsymbol{\theta} \in \mathbb{R}^d} \frac{1}{m} \sum_{i=1}^m \rho(Y_{i,m}^*, \hat{g}_m^{*(i)}(\mathbf{X}_{i,m}^{*\top} \boldsymbol{\beta}(\boldsymbol{\theta}) | \boldsymbol{\theta})) \hat{\tau}_{mi}^*(\boldsymbol{\theta})$$

and

$$\hat{\tau}_{mi}^*(\boldsymbol{\theta}) \equiv \tau_m\left(\hat{f}_m^{*(i)}(\mathbf{X}_{i,m}^{*\top} \boldsymbol{\beta}(\boldsymbol{\theta}) | \boldsymbol{\theta})\right),$$

where  $\tau_m(\cdot)$  is a smooth function on  $\mathbb{R}$  satisfying the conditions of Assumption 6 above.

Let  $\hat{g}_n^{(i,m)}(\mathbf{X}_{i,m}^{*\top} \boldsymbol{\beta}(\boldsymbol{\theta}) | \boldsymbol{\theta})$  and  $\hat{f}_n^{(i,m)}(\mathbf{X}_{i,m}^{*\top} \boldsymbol{\beta}(\boldsymbol{\theta}) | \boldsymbol{\theta})$  be the estimates of  $g_0$  and  $f_0$  given above in (2) and (3), respectively, that are constructed by summation over the elements in the original sample  $\mathcal{X}_n$  but excluding the observation appearing in the resample  $\mathcal{X}_m^*$  with index number  $i$ .

The following result shows how the particular resampling scheme adopted here may be used to reproduce the higher-order behaviour of  $\boldsymbol{\theta}_n$  in the bootstrap domain.

**Theorem 3** (Higher-Order Asymptotics for the Bootstrapped Estimator). *Recall the definition of  $\hat{\theta}_n$  as given above in (4), and recall the definitions of  $\gamma_1$  and  $\gamma_2$  given in the statement of Theorem 2 above. In addition, let  $\mathbf{A}_{m0}$  denote the quantity  $\mathbf{A}_{n0}$  given in the statement of Theorem 2 for a sample of size  $n = m$ . Suppose  $h_m \in [\delta m^{-\frac{1}{3}}, \delta^{-1} m^{-\frac{1}{3}}]$  for some small  $\delta \in (0, 1)$ .*

*Then under the conditions of Assumptions 1–6, we have*

1.

$$\sqrt{m} \left( \hat{\theta}_m^* - \hat{\theta}_n \right) = \mathbf{A}_{m0} + h_m^2 \sqrt{m} \gamma_1 + \frac{1}{h_m \sqrt{m}} \gamma_2 + o \left( m^{-\frac{1}{5}} \sqrt{\log m} \right)$$

*almost surely iff  $m \in [C_1 n^\xi, C_2 n^{1-\xi}]$  for suitable positive constants  $C_1, C_2$  and  $\xi$ ; while*

2. *if  $m = n$ , then*

$$\sqrt{m} \left( \hat{\theta}_m^* - \hat{\theta}_n \right) = \mathbf{A}_{m0} + \frac{1}{h_m \sqrt{m}} \gamma_2 + o \left( m^{-\frac{1}{5}} \sqrt{\log m} \right).$$

*Proof.* Appendix A.4. □

Let  $\hat{\theta}_m$  denote the original estimator given above in (4) of the population parameter  $\theta_0$  constructed from a sample of size  $m$ . Theorem 3 shows that it is possible to induce

$$\sqrt{m} \left( \hat{\theta}_m^* - \hat{\theta}_n \right) \stackrel{a.s.}{=} \sqrt{m} \left( \hat{\theta}_m - \theta_0 \right)$$

for large  $m$  and  $n$  up to higher-order terms of magnitude  $m^{-\frac{1}{6}}$  provided that the resample size  $m$  is sufficiently small relative to the original sample size  $n$ .<sup>13</sup> In

---

<sup>13</sup>The conclusion of Theorem 3 also illustrates the scope of what explicit methods of bias correction might be able to accomplish in this setting when combined with resampling. In particular, a method of correction for “degrees of freedom bias” that eliminates the term of order  $\frac{1}{h_n \sqrt{n}}$  in the expansion of Theorem 2 may be combined with a scheme of resampling  $m < n$  observations with replacement from the original sample of size  $n$  if one wishes to design a procedure to select a bandwidth for the resulting bias-corrected estimator. Combining an analytical correction for degrees of freedom bias in this context with the  $m$ -out-of- $n$  bootstrap would not be in essence different from the bandwidth selection procedure for the non-bias corrected estimator described below. On the other hand, a resampling scheme involving the generation of resamples of the same size as the original sample must be combined with a method of eliminating the bias term of order  $h_n^2 \sqrt{n}$  in the expansion of Theorem 2 in order to be of any use in bandwidth selection.

particular, it is immediate from the conclusion of Theorem 3 that the value of  $h_m$  that minimizes the order of the stochastic expansion of  $\sqrt{m} \left( \hat{\boldsymbol{\theta}}_m^* - \hat{\boldsymbol{\theta}}_n \right)$  is asymptotically equal with probability one to that which minimizes the stochastic expansion of  $\sqrt{m} \left( \hat{\boldsymbol{\theta}}_m - \boldsymbol{\theta}_0 \right)$ . In particular, if  $h_{m,opt}^*$  minimizes

$$E \left[ m \left( \hat{\boldsymbol{\theta}}_m^* - \hat{\boldsymbol{\theta}}_n \right)^\top \boldsymbol{\Sigma}_0^- \left( \hat{\boldsymbol{\theta}}_m^* - \hat{\boldsymbol{\theta}}_n \right) \right],$$

and if  $h_{m,opt}$  minimizes

$$E \left[ m \left( \hat{\boldsymbol{\theta}}_m - \boldsymbol{\theta}_0 \right)^\top \boldsymbol{\Sigma}_0^- \left( \hat{\boldsymbol{\theta}}_m - \boldsymbol{\theta}_0 \right) \right]$$

for  $\hat{\boldsymbol{\theta}}_m$  computed using a sample size of  $n = m$ , then

$$\frac{h_{m,opt}^*}{h_{m,opt}} \xrightarrow{a.s.} 1.$$

From this we have that

$$\hat{h}_{n,opt} \equiv h_{m,opt}^* \left( \frac{m}{n} \right)^{\frac{1}{3}} \quad (8)$$

is asymptotically equal almost surely to  $h_{n,opt}$  given above in (7).

In practice, one would generate  $B$  independent bootstrap resamples  $\mathcal{X}_{m,b}^*$  of size  $m$  (where  $b = \{1, \dots, B\}$ ) and choose  $h_m$  to minimize a Monte Carlo approximation to the bootstrap estimate  $E \left[ m \left( \hat{\boldsymbol{\theta}}_m^* - \hat{\boldsymbol{\theta}}_n \right)^\top \boldsymbol{\Sigma}_0^- \left( \hat{\boldsymbol{\theta}}_m^* - \hat{\boldsymbol{\theta}}_n \right) \middle| \mathcal{X}_n \right]$  of  $E \left[ m \left( \hat{\boldsymbol{\theta}}_m - \boldsymbol{\theta}_0 \right)^\top \boldsymbol{\Sigma}_0^- \left( \hat{\boldsymbol{\theta}}_m - \boldsymbol{\theta}_0 \right) \right]$ . In particular, if  $\hat{\boldsymbol{\theta}}_{m,b}^*$  denotes the semiparametric  $M$ -estimator computed using the observations appearing in the  $b$ th bootstrap resample  $\mathcal{X}_{m,b}^*$ , then  $h_{m,opt}^*$  would be taken to be the bandwidth value that minimizes

$$\frac{m}{B} \sum_{b=1}^B \left( \hat{\boldsymbol{\theta}}_{m,b}^* - \hat{\boldsymbol{\theta}}_n \right)^\top \hat{\boldsymbol{\Sigma}}_{m,B}^{*-} \left( \hat{\boldsymbol{\theta}}_{m,b}^* - \hat{\boldsymbol{\theta}}_n \right), \quad (9)$$

where  $\hat{\boldsymbol{\Sigma}}_{m,B}^{*-}$  is the generalized inverse of the empirical variance matrix of

$$\sqrt{m} \left( \hat{\boldsymbol{\theta}}_{m,b}^* - \hat{\boldsymbol{\theta}}_n \right), \quad b = 1, \dots, B$$



taken over the  $B$  bootstrap resamples.<sup>14</sup>

It remains to provide some guidance on the practical choice of the resample size  $m$ . In this connection, note that choosing  $h_{m,opt}^*$  to minimize the quantity in (9) above is asymptotically equivalent for large  $n$  and  $B$  to choosing  $h_{m,opt}^*$  to minimize  $E \left[ m \left( \hat{\theta}_m^* - \hat{\theta}_n \right)^\top \Sigma_0^- \left( \hat{\theta}_m^* - \hat{\theta}_n \right) \middle| \mathcal{X}_n \right]$ . In this connection, it is possible to provide some guidance in choosing the value of  $m$  that minimizes the magnitude of the estimation error involved in working with the estimated criterion  $E \left[ m \left( \hat{\theta}_m^* - \hat{\theta}_n \right)^\top \Sigma_0^- \left( \hat{\theta}_m^* - \hat{\theta}_n \right) \middle| \mathcal{X}_n \right]$  rather than with its estimand  $E \left[ m \left( \hat{\theta}_m^* - \hat{\theta}_n \right)^\top \Sigma_0^- \left( \hat{\theta}_m^* - \hat{\theta}_n \right) \right]$ . We have the following result.

**Theorem 4.** Suppose  $h_m \in \left[ \delta m^{-\frac{1}{3}}, \delta^{-1} m^{-\frac{1}{3}} \right]$  for some small  $\delta \in (0, 1)$  and  $m \in [C_1 n^\xi, C_2 n^{1-\xi}]$  for suitable positive constants  $C_1, C_2$  and  $\xi$ . Then under the conditions of Assumptions 1-6, a setting of

$$m \propto \left( \frac{n}{\log n} \right)^{\frac{3}{4}}$$

is sufficient to minimize the order of the estimation error

$$E \left[ m \left( \hat{\theta}_m^* - \hat{\theta}_n \right)^\top \Sigma_0^- \left( \hat{\theta}_m^* - \hat{\theta}_n \right) \middle| \mathcal{X}_n \right] - E \left[ m \left( \hat{\theta}_m^* - \hat{\theta}_n \right)^\top \Sigma_0^- \left( \hat{\theta}_m^* - \hat{\theta}_n \right) \right].$$

*Proof.* Appendix A.5. □

As such, Theorem 4 prescribes an optimal setting for the resample size in the form

$$m_{k,opt} = k \left( \frac{n}{\log n} \right)^{\frac{3}{4}} \tag{10}$$

for some  $k > 0$ , which immediately invites the question of how one should in practice choose the proportionality constant  $k$ .

---

<sup>14</sup>This method naturally requires the preliminary computation of the estimator  $\hat{\theta}_n$  using all the observations in the original sample. As such, a pilot bandwidth  $h_{n,pil}$  is required in order to implement the computation of  $\hat{\theta}_n$  over the full sample. In practice, one might select  $h_{n,pil}$  using some sort of rule of thumb or perhaps via the procedures of Härdle et al. (1993) or Xia et al. (2009).

In this connection, suppose one has  $B$  independent bootstrap resamples of size  $m = m_{k,opt}$  for  $m_{k,opt}$  as given in (10). Let  $\tilde{\boldsymbol{\theta}}_{m,b}^*$  denote the semiparametric  $M$ -estimator computed using the observations in the  $b$ th bootstrap resample and implemented using a suitable preliminary bandwidth  $h_{m,pre}$ . Set  $\tilde{\boldsymbol{\Sigma}}_{m_{k,opt},b}^*$  to be the empirical variance matrix of  $\sqrt{m} \left( \tilde{\boldsymbol{\theta}}_{m,b}^* - \hat{\boldsymbol{\theta}}_n \right)$  taken over the  $B$  bootstrap resamples. Following the basic suggestion of Bickel and Sakov (2008), it may be advisable in this case to choose  $k$  to minimize  $\left\| \tilde{\boldsymbol{\Sigma}}_{m_{k,opt},B} \right\|$ .<sup>15</sup>

## 4 Numerical Evidence

This section presents the results of a series of modest simulation experiments designed to assess the finite-sample reliability of the method of bandwidth selection proposed in this paper. In particular, the method described above is applied to the problem of implementing the semiparametric least-squares procedure of Ichimura (1993), i.e., the estimator

$$\hat{\boldsymbol{\theta}}_n = \arg \min_{\boldsymbol{\theta} \in \mathbb{R}^d} \sum_{i=1}^n \left( Y_i - \hat{g}_n^{(i)} \left( \mathbf{X}_i^\top \boldsymbol{\beta}(\boldsymbol{\theta}) \mid \boldsymbol{\theta} \right) \right)^2 \hat{\tau}_{ni}(\boldsymbol{\theta}) \quad (11)$$

of the normalized vector of index coefficients  $\boldsymbol{\theta}_0$  appearing in the regression relationship

$$E[Y \mid \mathbf{X}] = G_0 \left( \mathbf{X}^\top \boldsymbol{\beta}(\boldsymbol{\theta}_0) \right), \quad (12)$$

where

$$\boldsymbol{\beta}(\boldsymbol{\theta}_0) \equiv \begin{pmatrix} 1 & \boldsymbol{\theta}_0^\top \end{pmatrix}^\top.$$

The link function  $G_0(\cdot)$  appearing in (12) is naturally unknown and is estimated using the Nadaraya-Watson estimator given by  $\hat{g}_n^{(i)}(\cdot \mid \boldsymbol{\theta})$  appearing above in (2).

---

<sup>15</sup>It may also be feasible to construct an iterative procedure that alternates between bandwidth selection by minimization of

$$\frac{m}{B} \sum_{b=1}^B \left( \tilde{\boldsymbol{\theta}}_{m,b}^* - \hat{\boldsymbol{\theta}}_n \right)^\top \tilde{\boldsymbol{\Sigma}}_{m,B}^{*-} \left( \tilde{\boldsymbol{\theta}}_{m,b}^* - \hat{\boldsymbol{\theta}}_n \right)$$

and the selection of  $k$  by minimization of  $\left\| \tilde{\boldsymbol{\Sigma}}_{m_{k,opt},B} \right\|$ . In general, the practical selection of the resample size  $m$  is the subject of ongoing research. See e.g., Bickel and Sakov (2008) and the references cited therein.

In the series of Monte Carlo experiments presented here, 500 simulated samples each of size  $n = 1000$  were drawn from a population of ordered triples  $(X_1, X_2, U)^\top$ , where  $X_1$  and  $U$  are standard normal and  $X_2$  is exponential with unit mean.  $X_1$ ,  $X_2$  and  $U$  were set to be mutually independent, and taken to determine realizations of a binary response variable  $Y$  given as

$$Y = 1 \left\{ \begin{bmatrix} X_1 & X_2 \end{bmatrix} \begin{bmatrix} 1 \\ \theta_0 \end{bmatrix} + U > 0 \right\},$$

where the parameter of interest is taken to be  $\theta_0 \equiv -1$ .

The trimming function appearing in (11) was set to be identically equal to unity in the simulations, i.e.,  $\hat{\tau}_{ni}(\boldsymbol{\theta}) \equiv 1$  for all  $n$ ,  $i$  and  $\boldsymbol{\theta}$ . A quartic kernel of second order was used to compute the Nadaraya-Watson estimator given above in (2), to wit,

$$K(u) = \frac{15}{16} (1 - u^2)^2 \cdot 1 \{u^2 \leq 1\}. \quad (13)$$

Figures 1 and 2 below summarize the behaviour of the resulting semiparametric least-squares estimator  $\hat{\theta}_n$  over a range of bandwidths used to implement the Nadaraya-Watson estimator in (2). In particular, Figure 1 involves a plot against bandwidth of numerical approximations to the square root of the ratio of the mean squared error of  $\hat{\theta}_n$  to its variance—the so-called “root relative mean squared error” (RRMSE)—over the 1000 Monte Carlo replications considered. Figure 2 presents the same information, but with the axes adjusted to a log-log scale.

In particular, the value of the bandwidth that minimizes the simulated RRMSE of  $\hat{\theta}_n$  was found to be

$$h_{n,opt} = h_{1000,opt} = .4785.$$

This is compared in Table 1 to a number of  $m$ -bootstrap estimates of the optimal bandwidth as well as the bandwidth  $h_{1000,Sil}$  obtained by application of the rule-of-thumb of Silverman (1986). In the context of the quartic kernel given in (13) used here, the Silverman rule of thumb yields a bandwidth of the form

$$h_{n,Sil} = 2.7799 s_n n^{-\frac{1}{5}}, \quad (14)$$

where  $s_n$  is an estimate of the standard deviation of the index  $\mathbf{X}^\top \boldsymbol{\beta}(\boldsymbol{\theta})$ . In the context of the simulations presented here,  $s_n$  was set to be the empirical standard deviation of  $\{\mathbf{X}_i^\top \mathbf{b}_{OLS} : i = 1, \dots, n\}$ , where  $\mathbf{b}_{OLS}$  is the vector of ordinary least squares coefficients from a regression of  $Y$  on  $\mathbf{X}$  with the first element normalized to unity.

With respect to the bootstrap bandwidth estimates, consider  $B = 99$  bootstrap resamples each of size  $m$  drawn with replacement from the original dataset and take a particular RRMSE analogue given by

$$\sqrt{\frac{\frac{1}{B} \sum_{b=1}^B \left( \hat{\theta}_{m,b}^*(h) - \hat{\theta}_n(\hat{h}_{n,Sil}) \right)^2}{\sigma_B^{*2}(h)}}, \quad (15)$$

where  $\hat{\theta}_{m,b}^*(h)$  denotes the semiparametric least-squares estimator of  $\theta_0$  implemented with bandwidth  $h$  and computed using observations contained in the  $b$ th bootstrap resample, and where

$$\sigma_B^{*2}(h) \equiv \frac{1}{B-1} \sum_{b=1}^B \left( \hat{\theta}_{m,b}^*(h) - \frac{1}{B} \sum_{b'=1}^B \hat{\theta}_{m,b'}^*(h) \right)^2,$$

i.e., the empirical variance of  $\hat{\theta}_{m,b}^*(h)$  over the set of  $B$  bootstrap resamples of size  $m$ . The quantity  $\hat{\theta}_n(\hat{h}_{n,Sil})$  in (15) denotes a “pilot” estimate of  $\theta_0$  computed using the observations in the full sample of size  $n$  and implemented using the Silverman rule-of-thumb bandwidth  $\hat{h}_{n,Sil}$  given above in (14).

The  $m$ -bootstrap bandwidth estimates indicated in Table 1 were obtained by rescaling the bandwidth minimizing the expression in (15) over a grid of 100 evenly spaced points in the interval  $[.0001, 1]$  by a factor of  $\left(\frac{m}{n}\right)^{\frac{1}{3}}$ , as prescribed above in (8). The resulting bandwidth estimates are denoted  $\hat{h}_{m,opt}$ . Four different settings of resample size  $m$  were used to compute the expression in (15). Following the guidance given in Section 4,  $m$  was set to

$$m = \lfloor km^* \rfloor \equiv \left\lfloor k \left( \frac{n}{\log n} \right)^{\frac{3}{4}} \right\rfloor$$

for

$$k = 1, \frac{3}{2}, 2, 5,$$

which for  $n = 1000$  resulted in the settings of

$$m = 41, 62, 83, 208,$$

respectively.<sup>16</sup> The means and standard deviations of  $\hat{h}_{\lfloor km^* \rfloor, opt}$  over the 1000 Monte Carlo replications considered are summarized in Table 1 for  $k = 1, \frac{3}{2}, 2, 5$ ,

---

<sup>16</sup>Note that the notation  $\lfloor u \rfloor$  denotes the largest integer that is less than or equal to  $u$ .

along with their root mean squared errors with respect to the “true” value of  $h_{1000,opt} = .4785$ . It is found that the estimate  $\hat{h}_{\lfloor \frac{3}{2}m^* \rfloor, opt}$  seems to perform best in terms of root mean squared error out of the four  $m$ -bootstrap bandwidth estimates considered, although not by an overwhelmingly large margin. The finite-sample performance of the rule-of-thumb smoothing rule  $\hat{h}_{n, Sil}$  given above in (14) is also indicated in Table 1, and is clearly dominated by the four bootstrap bandwidth estimates considered.

Whether this makes any practical difference in the context of the specific data-generating process considered here is the focus of Table 2. In particular, Table 2 summarizes the finite-sample behaviour over 1000 Monte Carlo replications of the semiparametric least-squares estimator  $\hat{\theta}_n$  implemented using the six different smoothing rules considered in Table 1. Given that all six smoothing rules have sampling distributions that concentrate in the region of Figure 1 where the RRMSE of  $\hat{\theta}_n$  is both relatively flat and minimized, it is perhaps unsurprising that the behaviour of  $\hat{\theta}_n$  does not vary dramatically depending on which of the six bandwidths considered in Table 1 is used to implement the estimator. That said, however, the bootstrap bandwidth estimate computed using a resample size of  $m = \left\lfloor \frac{3}{2} \left( \frac{n}{\log n} \right)^{\frac{3}{4}} \right\rfloor = 62$  leads  $\hat{\theta}_n$  to have a slightly smaller root mean squared error than would be the case if the estimator was implemented using any of the four other feasible bandwidths considered. This result is in agreement with the results presented in Table 1 regarding the statistical behaviour of the smoothing rules themselves. The relative insensitivity of the behaviour of the semiparametric least-squares estimator shown in the simulation experiment presented here to the particular estimate of the optimal bandwidth used to implement it should perhaps be of some comfort for empirical practice.

## 5 Conclusion

This paper has considered the practical problem of implementing a semiparametric optimization estimator of the vector of index coefficients in a generalized single-index regression model with an unknown link function. The unknown link function is treated as a nuisance parameter and is taken to be estimated using a nonparametric kernel regression procedure of the Nadaraya-Watson type, while the criterion function in which the nonparametric estimate of the infinite-dimensional nuisance parameter appears is assumed to possess a sufficiently high order of differentiability in a neighbourhood of the true regression function. A par-

ticular resampling method was proposed in this paper for estimating an asymptotically optimal bandwidth required to implement the nonparametric estimate of the link function from the point of view of minimizing the expectation of the Mahalanobis distance between the estimator and the true value of the Euclidean interest parameter. In particular, the method proposed here involves the so-called “ $m$ -out-of- $n$ ”, or  $m$ -bootstrap—i.e., it involves the generation of resamples with replacement containing strictly fewer observations than are contained in the original dataset. This method is shown to work in the sense that the resulting  $m$ -bootstrap estimate of the average Mahalanobis distance is so close in a uniform sense in large samples to the actual expected value of the Mahalanobis distance of a parameter estimate computed using a dataset of  $m$  observations that the bandwidth that minimizes one is asymptotically equal almost surely to the bandwidth that minimizes the other. An estimate of the asymptotically optimal bandwidth for the parameter estimate computed using the full sample of size  $n$  can then be obtained by an appropriate rescaling of the bandwidth found to minimize the  $m$ -bootstrap estimate of the mean squared error.

Specific guidance on the practical selection of the resample size  $m$  was also given. In particular, for resamples of size  $m \in [C_1 n^\zeta, C_2 n^{1-\zeta}]$  for suitable positive constants  $C_1$ ,  $C_2$  and  $\zeta$ , it was shown that a setting of  $m \propto \left(\frac{n}{\log n}\right)^{\frac{3}{4}}$  is sufficient to minimize the stochastic order of the estimation error committed by substituting a bootstrap estimate of the average Mahalanobis distance for its actual population expectation. Simulation evidence presented here for the case of a semiparametric least-squares estimator of a binary-choice model seems to indicate that the parameter estimates implemented using the  $m$ -bootstrap bandwidth estimator has a small-sample behaviour that is relatively insensitive to the choice of resample size. Further work on the relationship between the sampling behaviour of bandwidth estimators and the sampling behaviour of the semiparametric estimators in which they are embedded would appear to be fruitful from both a theoretical and an applied viewpoint.

## References

- Andrews, D. W. K. (1994) ‘Asymptotics for semiparametric econometric models via stochastic equicontinuity.’ *Econometrica* 62, 43–72
- Bickel, P. J., and A. Sakov (2008) ‘On the choice of  $m$  in the  $m$  out of  $n$  bootstrap and confidence bounds for extrema.’ *Statistica Sinica* 18, 967–985

- Carroll, R. J., J. Fan, I. Gijbels, and M. P. Wand (1997) ‘Generalized partially linear single-index models.’ *Journal of the American Statistical Association* 92, 477–489
- Härdle, W., and T. M. Stoker (1989) ‘Investigating smooth multiple regression by the method of average derivatives.’ *Journal of the American Statistical Association* 84, 986–995
- Härdle, W., P. Hall, and H. Ichimura (1993) ‘Optimal smoothing in single-index models.’ *Annals of Statistics* 21, 157–178
- Hirano, K., G. W. Imbens, and G. Ridder (2003) ‘Efficient estimation of average treatment effects using the estimated propensity score.’ *Econometrica* 71, 1161–1189
- Horowitz, J. L. (2009) *Semiparametric and Nonparametric Methods in Econometrics* (New York: Springer Science+Business Media)
- Ichimura, H. (1993) ‘Semiparametric least squares (SLS) and weighted SLS estimation of single-index models.’ *Journal of Econometrics* 58, 71–120
- Ichimura, H., and O. Linton (2005) ‘Asymptotic expansions for some semiparametric program evaluation estimators.’ In *Identification and Inference for Econometric Models: Essays in Honor of Thomas Rothenberg*, ed. D. W. K. Andrews and J. H. Stock (New York: Cambridge University Press) pp. 149–170
- Klein, R. W., and R. H. Spady (1993) ‘An efficient semiparametric estimator for binary response models.’ *Econometrica* 61, 387–421
- Li, K. C. (1991) ‘Sliced inverse regression for dimension reduction.’ *Journal of the American Statistical Association* 86, 316–342. With discussion.
- Newey, W. K. (1994) ‘The asymptotic variance of semiparametric estimators.’ *Econometrica* 62, 1349–1382
- Nishiyama, Y., and P. M. Robinson (2005) ‘The bootstrap and Edgeworth correction for semiparametric averaged derivatives.’ *Econometrica* 73, 903–948
- Powell, J. L., J. H. Stock, and T. M. Stoker (1989) ‘Semiparametric estimation of index coefficients.’ *Econometrica* 57, 1403–1430
- Robinson, P. M. (1988) ‘Root- $n$ -consistent semiparametric regression.’ *Econometrica* 56, 931–954
- Sherman, R. P. (1994) ‘ $U$ -processes in the analysis of a generalized semiparametric regression estimator.’ *Econometric Theory* 10, 372–395
- Silverman, B. W. (1986) *Density Estimation for Statistics and Data Analysis* (London: Chapman and Hall)
- Xia, Y., W. Härdle, and O. Linton (2009) ‘Optimal smoothing for a computationally and statistically efficient single index estimator.’ Department of Economics, London School of Economics.

# A Appendix

## A.1 Lemmas

Begin with a recent martingale convergence result of Xia et al. (2009).

**Lemma 1** (Xia, Härdle & Linton (2009, Lemma 6.1)). *Suppose  $\{G_{ni}(x) : i = 1, \dots, n\}$  is a martingale with respect to  $\mathcal{F}_i \equiv \sigma \{G_{nl}(x) : l \leq i\}$  with  $x \in \mathcal{X}$ , where  $\mathcal{X}$  is a compact region in a multidimensional space such that*

1.  $|G_{ni}(x)| < \xi_i$ , where the  $\xi_i$  are iid and  $\sup E [\xi_1^{2r}] < \infty$  for some  $r > 2$ ;
2.  $E [G_{nk}^2(x)] < a_n s(x)$  with  $\inf s(x) > 0$ ;
3.  $|G_{ni}(x) - G_{ni}(\tilde{x})| < n^{\alpha_1} |x - \tilde{x}| M_i$  almost surely for some  $\alpha_1 > 0$  where  $M_i$  ( $i = 1, 2, \dots$ ) are iid with  $E [M_1^2] < \infty$ .

Then if  $a_n = cn^{-\delta}$  with  $0 \leq \delta < 1 - \frac{2}{r}$ , then for every  $\alpha'_1 > 0$  we have

$$\sup_{|x| \leq n^{\alpha'_1}} \left| \frac{1}{n} s^{-\frac{1}{2}}(x) \sum_{i=1}^n G_{ni}(x) \right| = O \left( \sqrt{\frac{a_n \log n}{n}} \right)$$

almost surely.

Lemma 1 is required for the uniform convergence results given in Lemma 2 below. In this connection, define

$$\Lambda_n \equiv \{u : |u| < n^c, f_0(u|\boldsymbol{\theta}) > n^{-2\zeta}, \boldsymbol{\theta} \in \Theta_n\}, \quad (16)$$

where  $c > \frac{1}{3}$  and  $\zeta > 0$  satisfies the constraints of Assumption 6. For a random matrix  $\mathbf{A}_n$  depending on both  $u$  and  $\boldsymbol{\theta}$ , write  $\mathbf{A}_n = \bar{O}(a_n)$  or  $\mathbf{A}_n = \bar{o}(a_n)$  iff all elements in  $\mathbf{A}_n$  are  $O(a_n)$  or  $o(a_n)$  almost surely and uniformly for  $\boldsymbol{\theta} \in \Theta_n$  and  $u \in \Lambda_n$ .

**Lemma 2.** *The following convergences hold as  $n \rightarrow \infty$  given the validity of Assumptions 1–6:*

1.  $\hat{f}_n^{(i)}(u|\boldsymbol{\theta}) = f_0(u|\boldsymbol{\theta}) + \bar{O}\left(h_n^2 + \sqrt{\frac{\log n}{nh_n}}\right)$ ;
2.  $\hat{g}_n^{(i)}(u|\boldsymbol{\theta}) = g_0(u|\boldsymbol{\theta}) + \bar{O}\left(h_n^2 + \sqrt{\frac{\log n}{nh_n}}\right)$ .

*Proof.* 1. Let

$$\hat{\epsilon}_{nf}^{(i)}(u|\boldsymbol{\theta}) \equiv \hat{f}_n^{(i)}(u|\boldsymbol{\theta}) - E \left[ \hat{f}_n^{(i)}(u|\boldsymbol{\theta}) \right].$$

We have  $\text{Var} \left[ \hat{\epsilon}_{nf}^{(i)}(u|\boldsymbol{\theta}) \right] = O\left(\frac{1}{nh_n}\right)$ , so from Lemma 1

$$\hat{\epsilon}_{nf}^{(i)}(u|\boldsymbol{\theta}) = \bar{O}\left(\sqrt{\frac{\log n}{nh_n}}\right).$$



Now let  $W_1(\boldsymbol{\theta}) \equiv \mathbf{X}_1^\top \boldsymbol{\beta}(\boldsymbol{\theta})$ . We have

$$\begin{aligned}
E \left[ \hat{f}_n^{(i)}(u|\boldsymbol{\theta}) \right] &= \frac{1}{h_n} E \left[ K \left( \frac{u - W_1(\boldsymbol{\theta})}{h_n} \right) \right] \\
&= \frac{1}{h_n} \int K \left( \frac{u - w}{h_n} \right) f_0(w|\boldsymbol{\theta}) dw \\
&= \int K(z) f_0(u - h_n z|\boldsymbol{\theta}) dz \\
&= f_0(u|\boldsymbol{\theta}) + \frac{h_n^2}{2} f_0''(u|\boldsymbol{\theta}) \int s^2 K(s) ds + \bar{O}(h_n^4).
\end{aligned}$$

The desired conclusion follows.

2. First consider the convergence of

$$\frac{1}{(n-1)h_n} \sum_{j \neq i} Y_j K_{h_n}(u - \mathbf{X}_j^\top \boldsymbol{\beta}(\boldsymbol{\theta})) = \hat{g}_n^{(i)}(u|\boldsymbol{\theta}) \hat{f}_n^{(i)}(u|\boldsymbol{\theta}).$$

We have

$$Var \left[ \hat{g}_n^{(i)}(u|\boldsymbol{\theta}) \hat{f}_n^{(i)}(u|\boldsymbol{\theta}) - E \left[ \hat{g}_n^{(i)}(u|\boldsymbol{\theta}) \hat{f}_n^{(i)}(u|\boldsymbol{\theta}) \right] \right] = O \left( \frac{1}{nh_n} \right),$$

so an application of Lemma 1 yields the convergence

$$\hat{g}_n^{(i)}(u|\boldsymbol{\theta}) \hat{f}_n^{(i)}(u|\boldsymbol{\theta}) - E \left[ \hat{g}_n^{(i)}(u|\boldsymbol{\theta}) \hat{f}_n^{(i)}(u|\boldsymbol{\theta}) \right] = \bar{O} \left( \sqrt{\frac{\log n}{nh_n}} \right).$$

For  $W_1(\boldsymbol{\theta}) \equiv \mathbf{X}_1^\top \boldsymbol{\beta}(\boldsymbol{\theta})$ ,

$$\begin{aligned}
&E \left[ \hat{g}_n^{(i)}(u|\boldsymbol{\theta}) \hat{f}_n^{(i)}(u|\boldsymbol{\theta}) \right] \\
&= \frac{1}{h_n} E \left[ Y_1 K \left( \frac{u - W_1(\boldsymbol{\theta})}{h_n} \right) \right] \\
&= \frac{1}{h_n} E \left[ E[Y_1 | W_1(\boldsymbol{\theta})] K \left( \frac{u - W_1(\boldsymbol{\theta})}{h_n} \right) \right] \\
&= \frac{1}{h_n} \int g_0(w|\boldsymbol{\theta}) K \left( \frac{u - w}{h_n} \right) f_0(w|\boldsymbol{\theta}) dw \\
&= \int g_0(u - h_n z|\boldsymbol{\theta}) K(z) f_0(u - h_n z|\boldsymbol{\theta}) dz \\
&= g_0(u|\boldsymbol{\theta}) f_0(u|\boldsymbol{\theta}) \\
&\quad + \frac{h_n^2}{2} [g_0''(u|\boldsymbol{\theta}) f_0(u|\boldsymbol{\theta}) + 2g_0(u|\boldsymbol{\theta}) f_0'(u|\boldsymbol{\theta}) + g_0(u|\boldsymbol{\theta}) f_0''(u|\boldsymbol{\theta})] \int s^2 K(s) ds \\
&\quad + \bar{O}(h_n^4).
\end{aligned}$$

As such,

$$\hat{g}_n^{(i)}(u|\boldsymbol{\theta}) \hat{f}_n^{(i)}(u|\boldsymbol{\theta}) = g_0(u|\boldsymbol{\theta}) f_0(u|\boldsymbol{\theta}) + \bar{O}\left(h_n^2 + \sqrt{\frac{\log n}{nh_n}}\right).$$

Combining this with the conclusion of the first part of this lemma yields the result

$$\hat{g}_n^{(i)}(u|\boldsymbol{\theta}) = g_0(u|\boldsymbol{\theta}) + \bar{O}\left(h_n^2 + \sqrt{\frac{\log n}{nh_n}}\right).$$

□

The next preliminary result considers the asymptotic effects of trimming. In particular, recall the definition of  $\Lambda_n$  given above in (16), and define a similar set

$$\tilde{\Lambda}_n \equiv \{\mathbf{x} : f_0(\mathbf{x}^\top \boldsymbol{\beta}(\boldsymbol{\theta})|\boldsymbol{\theta}) > 2n^{-c}\}. \quad (17)$$

We have the following.

**Lemma 3.** *For  $i = 1, \dots, n$ , define  $W_i \equiv \mathbf{X}_i^\top \boldsymbol{\beta}(\boldsymbol{\theta})$ . Then summations over the sets*

$$\{\mathbf{X}_i : i = 1, \dots, n\},$$

$$\{W_i : W_i \in \Lambda_n, i = 1, \dots, n\}$$

and

$$\{\mathbf{X}_i : \mathbf{X}_i \in \tilde{\Lambda}_n, i = 1, \dots, n\}$$

are interchangeable in the sense of almost sure consistency.

*Proof.* From Assumption 2 we have for  $\boldsymbol{\theta} \in \Theta$  that

$$\begin{aligned} \sum_{n=1}^{\infty} P[\cup_{i=1}^n \{W_i \notin \Lambda_n\}] &\leq \sum_{n=1}^{\infty} nP[W_n \notin \Lambda_n] \\ &\leq \sum_{n=1}^{\infty} nP[|W_n| > n^c] \\ &< \sum_{n=1}^{\infty} n \cdot n^{-6c} E[|W_n|^6] \\ &< \infty \end{aligned}$$

for all  $c > \frac{1}{3}$ . By the Borel-Cantelli Lemma we have

$$P[\cap_{n=1}^{\infty} \cup_{i=1}^n \{W_i \notin \Lambda_n\}] = 0.$$

A similar argument leads to the conclusion that

$$P[\cap_{n=1}^{\infty} \cup_{i=1}^n \{\mathbf{X}_i \notin \tilde{\Lambda}_n\}] = 0.$$

The desired conclusion is immediate. □

The next result is required in the derivation of the higher-order asymptotic representation of  $\sqrt{n}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0)$ . In this connection, define

$$\hat{w}_{ij}(\boldsymbol{\theta}) \equiv \frac{K_{h_n}(\mathbf{X}_i^\top \boldsymbol{\beta}(\boldsymbol{\theta}) - \mathbf{X}_j^\top \boldsymbol{\beta}(\boldsymbol{\theta}))}{(n-1)h_n \hat{f}_n^{(i)}(\mathbf{X}_i^\top \boldsymbol{\beta}(\boldsymbol{\theta})|\boldsymbol{\theta})} \quad (18)$$

and

$$w_{ij}(\boldsymbol{\theta}) \equiv \frac{K_{h_n}(\mathbf{X}_i^\top \boldsymbol{\beta}(\boldsymbol{\theta}) - \mathbf{X}_j^\top \boldsymbol{\beta}(\boldsymbol{\theta}))}{(n-1)h_n f_0(\mathbf{X}_i^\top \boldsymbol{\beta}(\boldsymbol{\theta})|\boldsymbol{\theta})}. \quad (19)$$

We have the following.

**Lemma 4.** 1.  $\hat{g}_n^{(i)}(\mathbf{X}_i^\top \boldsymbol{\beta}(\boldsymbol{\theta})|\boldsymbol{\theta}) - g_0(\mathbf{X}_i^\top \boldsymbol{\beta}(\boldsymbol{\theta})|\boldsymbol{\theta}) = \sum_{j \neq i} \epsilon_j(\boldsymbol{\theta}) w_{ij}(\boldsymbol{\theta}) + b_n(\mathbf{Z}_i|\boldsymbol{\theta})$ ,  
where

$$\epsilon_j(\boldsymbol{\theta}) \equiv Y_j - g_0(\mathbf{X}_j^\top \boldsymbol{\beta}(\boldsymbol{\theta})|\boldsymbol{\theta})$$

and

$$b_n(\mathbf{Z}_i|\boldsymbol{\theta}) \equiv \sum_{j \neq i} Y_j (\hat{w}_{ij}(\boldsymbol{\theta}) - w_{ij}(\boldsymbol{\theta})) + \sum_{j \neq i} g_0(\mathbf{X}_j^\top \boldsymbol{\beta}(\boldsymbol{\theta})|\boldsymbol{\theta}) w_{ij}(\boldsymbol{\theta}) - g_0(\mathbf{X}_i^\top \boldsymbol{\beta}(\boldsymbol{\theta})|\boldsymbol{\theta}).$$

$$2. b_n(\mathbf{Z}_i|\boldsymbol{\theta}) = \bar{O}\left(h_n^2 + \sqrt{\frac{\log n}{nh_n}}\right).$$

3. For  $\mathbf{z} \equiv (\mathbf{x}^\top, y)^\top$  and bandwidth  $h_n$  satisfying the requirements of Assumption 6,

$$\begin{aligned} & h_n^{-2} b_n(\mathbf{Z}_i|\boldsymbol{\theta}) \\ \xrightarrow{a.s.} & \frac{1}{2} \left[ g_0''(\mathbf{x}^\top \boldsymbol{\beta}(\boldsymbol{\theta})|\boldsymbol{\theta}) + \frac{2g_0'(\mathbf{x}^\top \boldsymbol{\beta}(\boldsymbol{\theta})|\boldsymbol{\theta}) f_0'(\mathbf{x}^\top \boldsymbol{\beta}(\boldsymbol{\theta})|\boldsymbol{\theta})}{f_0(\mathbf{x}^\top \boldsymbol{\beta}(\boldsymbol{\theta})|\boldsymbol{\theta})} + \frac{g_0(\mathbf{x}^\top \boldsymbol{\beta}(\boldsymbol{\theta})|\boldsymbol{\theta}) f_0''(\mathbf{x}^\top \boldsymbol{\beta}(\boldsymbol{\theta})|\boldsymbol{\theta})}{f_0(\mathbf{x}^\top \boldsymbol{\beta}(\boldsymbol{\theta})|\boldsymbol{\theta})} \right] \\ & \cdot \int s^2 K(s) ds \\ \equiv & b_0(\mathbf{z}|\boldsymbol{\theta}) \end{aligned}$$

uniformly for  $\mathbf{x} \in \tilde{\Lambda}_n$  and  $\boldsymbol{\theta} \in \Theta_n$ .

*Proof.* 1. Write

$$\begin{aligned} & \hat{g}_n^{(i)}(\mathbf{X}_i^\top \boldsymbol{\beta}(\boldsymbol{\theta})|\boldsymbol{\theta}) - g_0(\mathbf{X}_i^\top \boldsymbol{\beta}(\boldsymbol{\theta})|\boldsymbol{\theta}) \\ = & \sum_{j \neq i} Y_j \hat{w}_{ij}(\boldsymbol{\theta}) - g_0(\mathbf{X}_i^\top \boldsymbol{\beta}(\boldsymbol{\theta})|\boldsymbol{\theta}) \\ = & \sum_{j \neq i} Y_j (\hat{w}_{ij}(\boldsymbol{\theta}) - w_{ij}(\boldsymbol{\theta})) + \sum_{j \neq i} Y_j w_{ij}(\boldsymbol{\theta}) - g_0(\mathbf{X}_i^\top \boldsymbol{\beta}(\boldsymbol{\theta})|\boldsymbol{\theta}) + \sum_{j \neq i} g_0(\mathbf{X}_j^\top \boldsymbol{\beta}(\boldsymbol{\theta})|\boldsymbol{\theta}) w_{ij}(\boldsymbol{\theta}) \\ & - \sum_{j \neq i} g_0(\mathbf{X}_j^\top \boldsymbol{\beta}(\boldsymbol{\theta})|\boldsymbol{\theta}) w_{ij}(\boldsymbol{\theta}) \end{aligned}$$

and the conclusion is immediate.

2. Write  $b_n(\mathbf{Z}_i|\boldsymbol{\theta}) = b_{ni1}(\boldsymbol{\theta}) + b_{ni2}(\boldsymbol{\theta})$ , where

$$\begin{aligned} b_{ni1}(\boldsymbol{\theta}) &\equiv \sum_{j \neq i} Y_j (\hat{w}_{ij}(\boldsymbol{\theta}) - w_{ij}(\boldsymbol{\theta})); \\ b_{ni2}(\boldsymbol{\theta}) &\equiv \sum_{j \neq i} g_0(\mathbf{X}_j^\top \boldsymbol{\beta}(\boldsymbol{\theta})|\boldsymbol{\theta}) w_{ij}(\boldsymbol{\theta}) - g_0(\mathbf{X}_i^\top \boldsymbol{\beta}(\boldsymbol{\theta})|\boldsymbol{\theta}). \end{aligned}$$

Conditional on  $\mathbf{X}_i \in \tilde{\Lambda}_n$  and  $\boldsymbol{\theta} \in \Theta_n$  we have

$$\begin{aligned} &b_{ni1}(\boldsymbol{\theta}) \\ = &\frac{1}{\hat{f}_n^{(i)}(\mathbf{X}_i^\top \boldsymbol{\beta}(\boldsymbol{\theta})|\boldsymbol{\theta})} \\ &\cdot \left[ \frac{1}{(n-1)h_n} \sum_{j \neq i} Y_j K_{h_n}(\mathbf{X}_i^\top \boldsymbol{\beta}(\boldsymbol{\theta}) - \mathbf{X}_j^\top \boldsymbol{\beta}(\boldsymbol{\theta})) - \frac{1}{h_n} E[Y_1 K_{h_n}(\mathbf{X}_i^\top \boldsymbol{\beta}(\boldsymbol{\theta}) - \mathbf{X}_1^\top \boldsymbol{\beta}(\boldsymbol{\theta}))] \right] \\ &+ \frac{1}{\hat{f}_n^{(i)}(\mathbf{X}_i^\top \boldsymbol{\beta}(\boldsymbol{\theta})|\boldsymbol{\theta})} \cdot \frac{1}{h_n} E[Y_1 K_{h_n}(\mathbf{X}_i^\top \boldsymbol{\beta}(\boldsymbol{\theta}) - \mathbf{X}_1^\top \boldsymbol{\beta}(\boldsymbol{\theta}))] \\ &- \frac{1}{f_0(\mathbf{X}_i^\top \boldsymbol{\beta}(\boldsymbol{\theta})|\boldsymbol{\theta})} \\ &\cdot \left[ \frac{1}{(n-1)h_n} \sum_{j \neq i} Y_j K_{h_n}(\mathbf{X}_i^\top \boldsymbol{\beta}(\boldsymbol{\theta}) - \mathbf{X}_j^\top \boldsymbol{\beta}(\boldsymbol{\theta})) - \frac{1}{h_n} E[Y_1 K_{h_n}(\mathbf{X}_i^\top \boldsymbol{\beta}(\boldsymbol{\theta}) - \mathbf{X}_1^\top \boldsymbol{\beta}(\boldsymbol{\theta}))] \right] \\ &- \frac{1}{f_0(\mathbf{X}_i^\top \boldsymbol{\beta}(\boldsymbol{\theta})|\boldsymbol{\theta})} \cdot \frac{1}{h_n} E[Y_1 K_{h_n}(\mathbf{X}_i^\top \boldsymbol{\beta}(\boldsymbol{\theta}) - \mathbf{X}_1^\top \boldsymbol{\beta}(\boldsymbol{\theta}))] \\ = &\bar{O}\left(\sqrt{\frac{\log n}{nh_n}}\right) + g_0(\mathbf{X}_i^\top \boldsymbol{\beta}(\boldsymbol{\theta})|\boldsymbol{\theta}) + \bar{O}(h_n^2) + \bar{O}\left(\sqrt{\frac{\log n}{nh_n}}\right) - g_0(\mathbf{X}_i^\top \boldsymbol{\beta}(\boldsymbol{\theta})|\boldsymbol{\theta}) + \bar{O}(h_n^2) \\ = &\bar{O}\left(h_n^2 + \sqrt{\frac{\log n}{nh_n}}\right) \end{aligned}$$

by Lemmas 1 and 2 and a Taylor expansion of  $\frac{1}{h_n} E[Y_1 K_{h_n}(\mathbf{X}_i^\top \boldsymbol{\beta}(\boldsymbol{\theta}) - \mathbf{X}_1^\top \boldsymbol{\beta}(\boldsymbol{\theta}))]$ .

Now consider  $b_{ni2}(\boldsymbol{\theta})$ . Conditional on  $\mathbf{X}_i \in \tilde{\Lambda}_n$  and  $\boldsymbol{\theta} \in \Theta_n$  we have

$$\begin{aligned} &b_{ni2}(\boldsymbol{\theta}) \\ = &\frac{1}{f_0(\mathbf{X}_i^\top \boldsymbol{\beta}(\boldsymbol{\theta})|\boldsymbol{\theta})} \\ &\cdot \left[ \frac{1}{(n-1)h_n} \sum_{j \neq i} g_0(\mathbf{X}_j^\top \boldsymbol{\beta}(\boldsymbol{\theta})|\boldsymbol{\theta}) K_{h_n}(\mathbf{X}_i^\top \boldsymbol{\beta}(\boldsymbol{\theta}) - \mathbf{X}_j^\top \boldsymbol{\beta}(\boldsymbol{\theta})) \right. \\ &\left. - \frac{1}{h_n} E[g_0(\mathbf{X}_1^\top \boldsymbol{\beta}(\boldsymbol{\theta})|\boldsymbol{\theta}) K_{h_n}(\mathbf{X}_i^\top \boldsymbol{\beta}(\boldsymbol{\theta}) - \mathbf{X}_1^\top \boldsymbol{\beta}(\boldsymbol{\theta}))] \right] \end{aligned}$$

$$\begin{aligned}
& + \frac{1}{f_0(\mathbf{X}_i^\top \boldsymbol{\beta}(\boldsymbol{\theta}) | \boldsymbol{\theta})} \cdot \frac{1}{h_n} E [g_0(\mathbf{X}_1^\top \boldsymbol{\beta}(\boldsymbol{\theta}) | \boldsymbol{\theta}) K_{h_n}(\mathbf{X}_i^\top \boldsymbol{\beta}(\boldsymbol{\theta}) - \mathbf{X}_1^\top \boldsymbol{\beta}(\boldsymbol{\theta}))] \\
& - g_0(\mathbf{X}_i^\top \boldsymbol{\beta}(\boldsymbol{\theta}) | \boldsymbol{\theta}) \\
& = \bar{O} \left( \sqrt{\frac{\log n}{nh_n}} \right) + g_0(\mathbf{X}_i^\top \boldsymbol{\beta}(\boldsymbol{\theta}) | \boldsymbol{\theta}) + \bar{O}(h_n^2) - g_0(\mathbf{X}_i^\top \boldsymbol{\beta}(\boldsymbol{\theta}) | \boldsymbol{\theta}) \\
& = \bar{O} \left( h_n^2 + \sqrt{\frac{\log n}{nh_n}} \right)
\end{aligned}$$

by Lemma 1 and a Taylor expansion of  $\frac{1}{h_n} E [g_0(\mathbf{X}_1^\top \boldsymbol{\beta}(\boldsymbol{\theta}) | \boldsymbol{\theta}) K_{h_n}(\mathbf{X}_i^\top \boldsymbol{\beta}(\boldsymbol{\theta}) - \mathbf{X}_1^\top \boldsymbol{\beta}(\boldsymbol{\theta}))]$ .

It follows that  $b_n(\mathbf{Z}_i | \boldsymbol{\theta}) = \bar{O} \left( h_n^2 + \sqrt{\frac{\log n}{nh_n}} \right)$ .

3. Part 3 of this lemma is immediate from a Taylor expansion of

$$\frac{1}{h_n} E [g_0(\mathbf{X}_1^\top \boldsymbol{\beta}(\boldsymbol{\theta}) | \boldsymbol{\theta}) K_{h_n}(\mathbf{X}_i^\top \boldsymbol{\beta}(\boldsymbol{\theta}) - \mathbf{X}_1^\top \boldsymbol{\beta}(\boldsymbol{\theta}))].$$

□

The final result in this appendix is needed in the derivation of the stochastic expansion of the normalized bootstrapped estimator  $\sqrt{m}(\hat{\boldsymbol{\theta}}_m^* - \hat{\boldsymbol{\theta}}_n)$ . In this connection define

$$\hat{w}_{ij}^*(\boldsymbol{\theta}) \equiv \frac{K_{h_m}(\mathbf{X}_{i,m}^{*\top} \boldsymbol{\beta}(\boldsymbol{\theta}) - \mathbf{X}_{j,m}^{*\top} \boldsymbol{\beta}(\boldsymbol{\theta}))}{(m-1)h_m \hat{f}_m^{*(i)}(\mathbf{X}_{i,m}^{*\top} \boldsymbol{\beta}(\boldsymbol{\theta}) | \boldsymbol{\theta})} \quad (20)$$

and

$$w_{ij}^*(\boldsymbol{\theta}) \equiv \frac{K_{h_m}(\mathbf{X}_{i,m}^{*\top} \boldsymbol{\beta}(\boldsymbol{\theta}) - \mathbf{X}_{j,m}^{*\top} \boldsymbol{\beta}(\boldsymbol{\theta}))}{(m-1)h_m \hat{f}_n^{(i,m)}(\mathbf{X}_{i,m}^{*\top} \boldsymbol{\beta}(\boldsymbol{\theta}) | \boldsymbol{\theta})}. \quad (21)$$

For a random matrix  $\mathbf{A}_m^*$  depending on both  $u$  and  $\boldsymbol{\theta}$ , write  $\mathbf{A}_m^* = \bar{O}^*(a_m)$  or  $\mathbf{A}_m^* = \bar{o}^*(a_m)$  iff each element of  $\mathbf{A}_m^*$  is  $O(a_m)$  or  $o(a_m)$  almost surely and uniformly for  $u \in \Lambda_m$  and  $\boldsymbol{\theta} \in \Theta_n$  conditionally on  $\mathcal{X}_n$  as  $m, n \rightarrow \infty$ .<sup>17</sup> We have the following result analogous to Lemma 4.

**Lemma 5.** *1.*

$$\hat{g}_m^{*(i)}(\mathbf{X}_{i,m}^{*\top} \boldsymbol{\beta}(\boldsymbol{\theta}) | \boldsymbol{\theta}) - \hat{g}_n^{(i,m)}(\mathbf{X}_{i,m}^{*\top} \boldsymbol{\beta}(\boldsymbol{\theta}) | \boldsymbol{\theta}) = \sum_{j \neq m} \epsilon_{j,m}^*(\boldsymbol{\theta}) w_{ij}^*(\boldsymbol{\theta}) + b_m^*(\mathbf{Z}_{i,m}^* | \boldsymbol{\theta}),$$

where

$$\begin{aligned}
\epsilon_{j,m}^*(\boldsymbol{\theta}) & \equiv Y_{j,m}^* - \hat{g}_n^{(j,m)}(\mathbf{X}_{j,m}^{*\top} \boldsymbol{\beta}(\boldsymbol{\theta}) | \boldsymbol{\theta}); \\
b_m^*(\mathbf{Z}_{i,m}^* | \boldsymbol{\theta}) & \equiv \sum_{j \neq i} Y_{j,m}^* (\hat{w}_{ij}^*(\boldsymbol{\theta}) - w_{ij}^*(\boldsymbol{\theta})) \\
& \quad + \sum_{j \neq i} \hat{g}_n^{(j,m)}(\mathbf{X}_{j,m}^{*\top} \boldsymbol{\beta}(\boldsymbol{\theta}) | \boldsymbol{\theta}) w_{ij}^*(\boldsymbol{\theta}) - \hat{g}_n^{(i,m)}(\mathbf{X}_{i,m}^{*\top} \boldsymbol{\beta}(\boldsymbol{\theta}) | \boldsymbol{\theta}).
\end{aligned}$$

<sup>17</sup>Here  $\Lambda_m$  is as given above in (16) with  $n = m$ .

2.  $E \left[ b_m^* (\mathbf{Z}_{i,m}^* | \boldsymbol{\theta}) | \mathcal{X}_n, \mathbf{X}_{i,m}^* \in \tilde{\Lambda}_m \right] = \frac{1}{n} \sum_i b_m^* (\mathbf{Z}_i | \boldsymbol{\theta}) = \bar{O} \left( h_n^2 + \sqrt{\frac{\log n}{nh_n}} \right)$ , where  $\tilde{\Lambda}_m$  is as given above in (17) for  $n = m$ .
3. (a) If  $m \in [C_1 n^\xi, C_2 n^{1-\xi}]$  for suitable positive constants  $C_1, C_2$  and  $\xi$  and  $h_m$  satisfies the requirements of Assumption 6 for  $n = m$ , then

$$\begin{aligned}
& h_m^{-2} E \left[ b_m^* (\mathbf{Z}_{i,m}^* | \boldsymbol{\theta}) | \mathcal{X}_n, \mathbf{X}_{i,m}^* \in \tilde{\Lambda}_m \right] \\
& \xrightarrow{a.s.} \frac{1}{2} \left[ g_0'' (\mathbf{X}_{i,m}^{*\top} \boldsymbol{\beta}(\boldsymbol{\theta}) | \boldsymbol{\theta}) + \frac{2g_0'' (\mathbf{X}_{i,m}^{*\top} \boldsymbol{\beta}(\boldsymbol{\theta}) | \boldsymbol{\theta}) f_0' (\mathbf{X}_{i,m}^{*\top} \boldsymbol{\beta}(\boldsymbol{\theta}) | \boldsymbol{\theta})}{f_0 (\mathbf{X}_{i,m}^{*\top} \boldsymbol{\beta}(\boldsymbol{\theta}) | \boldsymbol{\theta})} \right. \\
& \quad \left. + \frac{g_0 (\mathbf{X}_{i,m}^{*\top} \boldsymbol{\beta}(\boldsymbol{\theta}) | \boldsymbol{\theta}) f_0'' (\mathbf{X}_{i,m}^{*\top} \boldsymbol{\beta}(\boldsymbol{\theta}) | \boldsymbol{\theta})}{f_0 (\mathbf{X}_{i,m}^{*\top} \boldsymbol{\beta}(\boldsymbol{\theta}) | \boldsymbol{\theta})} \right] \int s^2 K(s) ds \\
& = b_0 ((\mathbf{X}_{i,m}^{*\top}, Y_i)^\top | \boldsymbol{\theta})
\end{aligned}$$

uniformly for  $\mathbf{X}_{i,m}^* \in \tilde{\Lambda}_m$  and  $\boldsymbol{\theta} \in \Theta_n$  and where  $b_0 ((\mathbf{X}_{i,m}^{*\top}, Y_i)^\top | \boldsymbol{\theta})$  is as in the statement of Lemma 4 above.

- (b) If  $m = n$ , then  $h_m^{-2} E \left[ b_m^* (\mathbf{Z}_{i,m}^* | \boldsymbol{\theta}) | \mathcal{X}_n, \mathbf{X}_{i,m}^* \in \tilde{\Lambda}_m \right]$  diverges.

*Proof.* 1. Write

$$\begin{aligned}
& \hat{g}_m^{*(i)} (\mathbf{X}_{i,m}^{*\top} \boldsymbol{\beta}(\boldsymbol{\theta}) | \boldsymbol{\theta}) - \hat{g}_n^{(i,m)} (\mathbf{X}_{i,m}^{*\top} \boldsymbol{\beta}(\boldsymbol{\theta}) | \boldsymbol{\theta}) \\
& = \sum_{j \neq i} Y_{j,m}^* \hat{w}_{ij}^* (\boldsymbol{\theta}) - \hat{g}_n^{(i,m)} (\mathbf{X}_{i,m}^{*\top} \boldsymbol{\beta}(\boldsymbol{\theta}) | \boldsymbol{\theta}) \\
& = \sum_{j \neq i} Y_{j,m}^* (\hat{w}_{ij}^* (\boldsymbol{\theta}) - w_{ij}^* (\boldsymbol{\theta})) + \sum_{j \neq i} Y_{j,m}^* w_{ij}^* (\boldsymbol{\theta}) - \hat{g}_n^{(i,m)} (\mathbf{X}_{i,m}^{*\top} \boldsymbol{\beta}(\boldsymbol{\theta}) | \boldsymbol{\theta}) \\
& \quad + \sum_{j \neq i} \hat{g}_n^{(j,m)} (\mathbf{X}_{j,m}^{*\top} \boldsymbol{\beta}(\boldsymbol{\theta}) | \boldsymbol{\theta}) w_{ij}^* (\boldsymbol{\theta}) - \sum_{j \neq i} \hat{g}_n^{(j,m)} (\mathbf{X}_{j,m}^{*\top} \boldsymbol{\beta}(\boldsymbol{\theta}) | \boldsymbol{\theta}) w_{ij}^* (\boldsymbol{\theta})
\end{aligned}$$

and the conclusion is immediate.

2. Write  $b_m^* (\mathbf{Z}_{i,m}^* | \boldsymbol{\theta}) = b_{mi1}^* (\boldsymbol{\theta}) + b_{mi2}^* (\boldsymbol{\theta})$ , where

$$b_{mi1}^* (\boldsymbol{\theta}) \equiv \sum_{j \neq i} Y_{j,m}^* (\hat{w}_{ij}^* (\boldsymbol{\theta}) - w_{ij}^* (\boldsymbol{\theta}))$$

and

$$b_{mi2}^* (\boldsymbol{\theta}) \equiv \sum_{j \neq i} \hat{g}_n^{(j,m)} (\mathbf{X}_{j,m}^{*\top} \boldsymbol{\beta}(\boldsymbol{\theta}) | \boldsymbol{\theta}) w_{ij}^* (\boldsymbol{\theta}) - \hat{g}_n^{(i,m)} (\mathbf{X}_{i,m}^{*\top} \boldsymbol{\beta}(\boldsymbol{\theta}) | \boldsymbol{\theta}).$$

Consider  $b_{mi1}^* (\boldsymbol{\theta})$ . We have

$$\begin{aligned}
& b_{mi1}^* (\boldsymbol{\theta}) \\
& = \frac{\hat{f}_n^{(i,m)} (\mathbf{X}_{i,m}^{*\top} \boldsymbol{\beta}(\boldsymbol{\theta}) | \boldsymbol{\theta}) - \hat{f}_m^{*(i)} (\mathbf{X}_{i,m}^{*\top} \boldsymbol{\beta}(\boldsymbol{\theta}) | \boldsymbol{\theta})}{\hat{f}_m^{*(i)} (\mathbf{X}_{i,m}^{*\top} \boldsymbol{\beta}(\boldsymbol{\theta}) | \boldsymbol{\theta}) \hat{f}_n^{(i,m)} (\mathbf{X}_{i,m}^{*\top} \boldsymbol{\beta}(\boldsymbol{\theta}) | \boldsymbol{\theta})} \\
& \quad \cdot \frac{1}{(m-1)h_m} \sum_{j \neq i} Y_{j,m}^* K_{h_m} (\mathbf{X}_{i,m}^{*\top} \boldsymbol{\beta}(\boldsymbol{\theta}) - \mathbf{X}_{j,m}^{*\top} \boldsymbol{\beta}(\boldsymbol{\theta})).
\end{aligned}$$

Note that

$$\begin{aligned}
& \hat{f}_m^{*(i)}(\mathbf{X}_{i,m}^{*\top} \boldsymbol{\beta}(\boldsymbol{\theta}) | \boldsymbol{\theta}) \\
&= \frac{1}{nh_m} \sum_{j=1}^n K_{h_m}(\mathbf{X}_{i,m}^{*\top} \boldsymbol{\beta}(\boldsymbol{\theta}) - \mathbf{X}_j^\top \boldsymbol{\beta}(\boldsymbol{\theta})) + \bar{O}^* \left( \sqrt{\frac{\log m}{mh_m}} \right) \\
&= \bar{O} \left( h_m^2 + \sqrt{\frac{\log n}{nh_m}} \right) + \bar{O}^* \left( \sqrt{\frac{\log m}{mh_m}} \right) \\
&= \bar{O} \left( h_n^2 + \sqrt{\frac{\log n}{nh_n}} \right),
\end{aligned}$$

where an appeal to Lemma 1 has been made. Combined with a further application of Lemma 1 we have

$$E \left[ b_{mi1}^*(\boldsymbol{\theta}) | \mathcal{X}_n, \mathbf{X}_{i,m}^* \in \tilde{\Lambda}_m \right] = \bar{O} \left( h_n^2 + \sqrt{\frac{\log n}{nh_n}} \right).$$

On the other hand,

$$\begin{aligned}
& E \left[ b_{mi2}^*(\boldsymbol{\theta}) | \mathcal{X}_n, \mathbf{X}_{i,m}^* \in \tilde{\Lambda}_m \right] \\
&= \frac{1}{n} \sum_{k=1}^n \hat{g}_n^{(k)}(\mathbf{X}_k^\top \boldsymbol{\beta}(\boldsymbol{\theta}) | \boldsymbol{\theta}) \cdot \frac{K_{h_m}(\mathbf{X}_{i,m}^{*\top} \boldsymbol{\beta}(\boldsymbol{\theta}) - \mathbf{X}_k^\top \boldsymbol{\beta}(\boldsymbol{\theta}))}{h_m \hat{f}_n^{(i,m)}(\mathbf{X}_{i,m}^{*\top} \boldsymbol{\beta}(\boldsymbol{\theta}) | \boldsymbol{\theta})} \\
&\quad - \hat{g}_n^{(i,m)}(\mathbf{X}_{i,m}^{*\top} \boldsymbol{\beta}(\boldsymbol{\theta}) | \boldsymbol{\theta}).
\end{aligned} \tag{22}$$

Note that when  $m \in [C_1 n^\xi, C_2 n^{1-\xi}]$  for suitable positive constants  $C_1, C_2$  and  $\xi$ ,

$$\begin{aligned}
& E \left[ b_{mi2}^*(\boldsymbol{\theta}) | \mathcal{X}_n, \mathbf{X}_{i,m}^* \in \tilde{\Lambda}_m \right] \\
&= \frac{E \left[ g_0(\mathbf{X}_1^\top \boldsymbol{\beta}(\boldsymbol{\theta}) | \boldsymbol{\theta}) K_{h_m}(\mathbf{X}_{i,m}^{*\top} \boldsymbol{\beta}(\boldsymbol{\theta}) - \mathbf{X}_1^\top \boldsymbol{\beta}(\boldsymbol{\theta})) \right]}{h_m f_0(\mathbf{X}_{i,m}^{*\top} \boldsymbol{\beta}(\boldsymbol{\theta}) | \boldsymbol{\theta})} + \bar{O} \left( \sqrt{\frac{\log n}{nh_m}} \right) \\
&\quad - g_0(\mathbf{X}_{i,m}^{*\top} \boldsymbol{\beta}(\boldsymbol{\theta}) | \boldsymbol{\theta}) + \bar{O} \left( h_n^2 + \sqrt{\frac{\log n}{nh_n}} \right) \\
&= \bar{O} \left( h_n^2 + \sqrt{\frac{\log n}{nh_n}} \right).
\end{aligned} \tag{23}$$

If, however,  $m = n$  we have

$$\begin{aligned}
& E \left[ b_{mi2}^*(\boldsymbol{\theta}) | \mathcal{X}_n, \mathbf{X}_{i,m}^* \in \tilde{\Lambda}_m \right] \\
&= \frac{1}{\hat{f}_n^{(i,m)}(\mathbf{X}_{i,m}^{*\top} \boldsymbol{\beta}(\boldsymbol{\theta}) | \boldsymbol{\theta})}
\end{aligned}$$

$$\begin{aligned}
& \cdot \frac{1}{nh_n} \sum_{k=1}^n \left( \hat{g}_n^{(k)} (\mathbf{X}_k^\top \boldsymbol{\beta}(\boldsymbol{\theta}) | \boldsymbol{\theta}) - Y_k \right) K_{h_n} (\mathbf{X}_{i,m}^{*\top} \boldsymbol{\beta}(\boldsymbol{\theta}) - \mathbf{X}_k^\top \boldsymbol{\beta}(\boldsymbol{\theta})) + \bar{o}(1) \\
& = \frac{1}{f_0 (\mathbf{X}_{i,m}^{*\top} \boldsymbol{\beta}(\boldsymbol{\theta}) | \boldsymbol{\theta})} \left( \frac{1}{h_n} E [\epsilon_1(\boldsymbol{\theta}) K_{h_n} (\mathbf{X}_{i,m}^{*\top} \boldsymbol{\beta}(\boldsymbol{\theta}) - \mathbf{X}_1^\top \boldsymbol{\beta}(\boldsymbol{\theta}))] + \bar{O} \left( \sqrt{\frac{\log n}{nh_n}} \right) \right) \\
& \quad + \bar{o}(1) \\
& = \bar{O} \left( \sqrt{\frac{\log n}{nh_n}} \right)
\end{aligned}$$

by virtue of  $E [\epsilon_1(\boldsymbol{\theta}) | \mathbf{X}_1] \stackrel{a.s.}{=} 0$  and via applications of Lemma 1.

3. (a) The desired convergence is a direct consequence of the presence of

$$\frac{1}{h_n} E [g_0 (\mathbf{X}_1^\top \boldsymbol{\beta}(\boldsymbol{\theta}) | \boldsymbol{\theta}) K_{h_m} (\mathbf{X}_{i,m}^{*\top} \boldsymbol{\beta}(\boldsymbol{\theta}) - \mathbf{X}_1^\top \boldsymbol{\beta}(\boldsymbol{\theta}))]$$

in (23) above. In particular, the desired conclusion can be obtained via a Taylor expansion of  $\frac{1}{h_n} E [g_0 (\mathbf{X}_1^\top \boldsymbol{\beta}(\boldsymbol{\theta}) | \boldsymbol{\theta}) K_{h_m} (\mathbf{X}_{i,m}^{*\top} \boldsymbol{\beta}(\boldsymbol{\theta}) - \mathbf{X}_1^\top \boldsymbol{\beta}(\boldsymbol{\theta}))]$ .

- (b) This conclusion follows from the absence of

$$\frac{1}{h_n} E [g_0 (\mathbf{X}_1^\top \boldsymbol{\beta}(\boldsymbol{\theta}) | \boldsymbol{\theta}) K_{h_m} (\mathbf{X}_{i,m}^{*\top} \boldsymbol{\beta}(\boldsymbol{\theta}) - \mathbf{X}_1^\top \boldsymbol{\beta}(\boldsymbol{\theta}))]$$

in the expression for  $E [b_{mi2}^*(\boldsymbol{\theta}) | \mathcal{X}_n, \mathbf{X}_{i,m}^* \in \tilde{\Lambda}_m]$  when  $m = n$ .

□

## A.2 Proof of Theorem 1

Note from Lemma 2 that

$$\hat{f}_n^{(i)}(u | \boldsymbol{\theta}) = f_0(u | \boldsymbol{\theta}) + \bar{O} \left( h_n^2 + \sqrt{\frac{\log n}{nh_n}} \right).$$

From the smoothness of the trimming function  $\tau_n(\cdot)$  we have

$$\hat{\tau}_{ni}(\boldsymbol{\theta}) = \tilde{\tau}_{ni}(\boldsymbol{\theta}) + \bar{O} \left( n^\zeta \left( h_n^2 + \sqrt{\frac{\log n}{nh_n}} \right) \right), \quad (24)$$

where

$$\tilde{\tau}_{ni}(\boldsymbol{\theta}) \equiv \tau_n(f_0(\mathbf{X}_i^\top \boldsymbol{\beta}(\boldsymbol{\theta}) | \boldsymbol{\theta})). \quad (25)$$

Let  $\bar{\sum}_i$  denote summations over indices  $i$  such that  $\mathbf{X}_i \in \tilde{\Lambda}_n$  for  $\tilde{\Lambda}_n$  as given above in (17). For  $\boldsymbol{\theta} \in \Theta_n$ , let

$$\boldsymbol{\gamma}_n(\boldsymbol{\theta}) \equiv \boldsymbol{\theta}_0 + \sqrt{n}(\boldsymbol{\theta} - \boldsymbol{\theta}_0),$$

and let

$$\tilde{\rho}(\mathbf{Z}_i, \boldsymbol{\theta}) \equiv \rho(Y_i, g_0(\mathbf{X}_i^\top \boldsymbol{\beta}(\boldsymbol{\theta}) | \boldsymbol{\theta})).$$



Abuse existing notation and let  $\hat{S}_n(\boldsymbol{\theta})$  denote  $\hat{S}_n(\gamma_n(\boldsymbol{\theta}))$ ,  $g_0(\mathbf{X}_i^\top \boldsymbol{\beta}(\boldsymbol{\theta}) | \boldsymbol{\theta})$  denote

$$g_0(\mathbf{X}_i^\top \boldsymbol{\beta}(\gamma_n(\boldsymbol{\theta})) | \gamma_n(\boldsymbol{\theta})),$$

$\hat{g}_n^{(i)}(\mathbf{X}_i^\top \boldsymbol{\beta}(\boldsymbol{\theta}) | \boldsymbol{\theta})$  denote  $\hat{g}_n^{(i)}(\mathbf{X}_i^\top \boldsymbol{\beta}(\gamma_n(\boldsymbol{\theta})) | \gamma_n(\boldsymbol{\theta}))$ ,  $\hat{\tau}_{ni}(\boldsymbol{\theta})$  denote  $\hat{\tau}_{ni}(\gamma_n(\boldsymbol{\theta}))$ , and so on *mutatis mutandis*.

Expand the function  $\rho(Y_i, \hat{g}_n^{(i)}(\mathbf{X}_i^\top \boldsymbol{\beta}(\boldsymbol{\theta}) | \boldsymbol{\theta}))$  about  $g_0(\mathbf{X}_i^\top \boldsymbol{\beta}(\boldsymbol{\theta}) | \boldsymbol{\theta})$  to third-order terms to arrive at

$$\begin{aligned} & \hat{S}_n(\boldsymbol{\theta}) \\ = & \frac{1}{n} \tilde{\rho}(\mathbf{Z}_i, \boldsymbol{\theta}) \hat{\tau}_{ni}(\boldsymbol{\theta}) \end{aligned} \quad (26)$$

$$\begin{aligned} & + \sum_{m=1}^2 \frac{1}{m!n} \sum_{i=1}^n \rho^{(m)}(Y_i, g_0(\mathbf{X}_i^\top \boldsymbol{\beta}(\boldsymbol{\theta}) | \boldsymbol{\theta})) \\ & \cdot \left( \hat{g}_n^{(i)}(\mathbf{X}_i^\top \boldsymbol{\beta}(\boldsymbol{\theta}) | \boldsymbol{\theta}) - g_0(\mathbf{X}_i^\top \boldsymbol{\beta}(\boldsymbol{\theta}) | \boldsymbol{\theta}) \right)^m \hat{\tau}_{ni}(\boldsymbol{\theta}) \end{aligned} \quad (27)$$

$$+ \frac{1}{3!n} \sum_{i=1}^n \rho^{(3)}(Y_i, u_{ni}(\boldsymbol{\theta})) \left( \hat{g}_n^{(i)}(\mathbf{X}_i^\top \boldsymbol{\beta}(\boldsymbol{\theta}) | \boldsymbol{\theta}) - g_0(\mathbf{X}_i^\top \boldsymbol{\beta}(\boldsymbol{\theta}) | \boldsymbol{\theta}) \right)^3 \hat{\tau}_{ni}(\boldsymbol{\theta}), \quad (28)$$

where  $u_{ni}(\boldsymbol{\theta})$  is between  $\hat{g}_n^{(i)}(\mathbf{X}_i^\top \boldsymbol{\beta}(\boldsymbol{\theta}) | \boldsymbol{\theta})$  and  $g_0(\mathbf{X}_i^\top \boldsymbol{\beta}(\boldsymbol{\theta}) | \boldsymbol{\theta})$ .

Consider (26). By (24) and appealing to Lemma 3 we have

$$\begin{aligned} & \frac{1}{n} \sum_{i=1}^n \tilde{\rho}(\mathbf{Z}_i, \boldsymbol{\theta}) \hat{\tau}_{ni}(\boldsymbol{\theta}) \\ = & \frac{1}{n} \sum_{i=1}^n \tilde{\rho}(\mathbf{Z}_i, \boldsymbol{\theta}) \tilde{\tau}_{ni}(\boldsymbol{\theta}) + \bar{O} \left( n^\zeta \left( h_n^2 + \sqrt{\frac{\log n}{nh_n}} \right) \right) \\ = & \frac{1}{n} \sum_i \tilde{\rho}(\mathbf{Z}_i, \boldsymbol{\theta}) + \bar{O} \left( n^\zeta \left( h_n^2 + \sqrt{\frac{\log n}{nh_n}} \right) \right) \\ = & \frac{1}{n} \sum_i \tilde{\rho}(\mathbf{Z}_i, \boldsymbol{\theta}) + \bar{o}(1) \end{aligned} \quad (29)$$

Expanding  $\tilde{\rho}(\mathbf{Z}_i, \boldsymbol{\theta})$  about  $\boldsymbol{\theta}_0$ , and recalling the abuse of notation that equates  $\tilde{\rho}(\mathbf{Z}_i, \boldsymbol{\theta})$  with  $\tilde{\rho}(\mathbf{Z}_i, \boldsymbol{\theta}_0 + \sqrt{n}(\boldsymbol{\theta} - \boldsymbol{\theta}_0))$  yields

$$\begin{aligned} & \frac{1}{n} \sum_i \tilde{\rho}(\mathbf{Z}_i, \boldsymbol{\theta}) \\ = & \frac{1}{n} \sum_i \tilde{\rho}(\mathbf{Z}_i, \boldsymbol{\theta}_0) + (\boldsymbol{\theta} - \boldsymbol{\theta}_0)^\top \frac{1}{\sqrt{n}} \sum_i \nabla_1 \tilde{\rho}(\mathbf{Z}_i, \boldsymbol{\theta}_0) \\ & + \frac{n}{2} (\boldsymbol{\theta} - \boldsymbol{\theta}_0)^\top \cdot \frac{1}{n} \sum_i \nabla_2 \tilde{\rho}(\mathbf{Z}_i, \boldsymbol{\theta}_0) (\boldsymbol{\theta} - \boldsymbol{\theta}_0) + \bar{o}(n \|\boldsymbol{\theta} - \boldsymbol{\theta}_0\|^2). \end{aligned} \quad (30)$$

Note that

$$\frac{1}{\sqrt{n}} \sum_i \nabla_1 \tilde{\rho}(\mathbf{Z}_i, \boldsymbol{\theta}_0) \xrightarrow{d} N(\mathbf{0}, \mathbf{D}_0),$$

where

$$D_0 \equiv E \left[ \nabla_1 \tilde{\rho}(\mathbf{Z}_1, \boldsymbol{\theta}_0) \nabla_1 \tilde{\rho}(\mathbf{Z}_1, \boldsymbol{\theta}_0)^\top \right].$$

Combining (29) and (30), the desired conclusion is immediate if it can be shown that both (27) and (28) are  $\bar{o}(1)$ .

Consider (27). By Lemma 2, we have

$$\left| \hat{g}_n^{(i)}(u|\boldsymbol{\theta}) - g_0(\mathbf{X}_i^\top \boldsymbol{\beta}(\boldsymbol{\theta})|\boldsymbol{\theta}) \right| = \bar{O} \left( h_n^2 + \sqrt{\frac{\log n}{nh_n}} \right),$$

while by Assumptions 3 and 6 we have

$$\begin{aligned} & \sum_{m=1}^2 \frac{1}{m!n} \sum_{i=1}^n \rho^{(m)}(Y_i, g_0(\mathbf{X}_i^\top \boldsymbol{\beta}(\boldsymbol{\theta})|\boldsymbol{\theta})) \left( \hat{g}_n^{(i)}(\mathbf{X}_i^\top \boldsymbol{\beta}(\boldsymbol{\theta})|\boldsymbol{\theta}) - g_0(\mathbf{X}_i^\top \boldsymbol{\beta}(\boldsymbol{\theta})|\boldsymbol{\theta}) \right)^m \hat{\tau}_{ni}(\boldsymbol{\theta}) \\ &= \sum_{m=1}^2 \frac{1}{m!n} \sum_{i=1}^n \rho^{(m)}(Y_i, g_0(\mathbf{X}_i^\top \boldsymbol{\beta}(\boldsymbol{\theta})|\boldsymbol{\theta})) \left( \hat{g}_n^{(i)}(\mathbf{X}_i^\top \boldsymbol{\beta}(\boldsymbol{\theta})|\boldsymbol{\theta}) - g_0(\mathbf{X}_i^\top \boldsymbol{\beta}(\boldsymbol{\theta})|\boldsymbol{\theta}) \right)^m \hat{\tau}_{ni}(\boldsymbol{\theta}) \\ & \quad + \bar{o} \left( n^\zeta \left( h_n^2 + \sqrt{\frac{\log n}{nh_n}} \right) \right) \\ &= \sum_{m=1}^2 \frac{1}{m!n} \sum_i \rho^{(m)}(Y_i, g_0(\mathbf{X}_i^\top \boldsymbol{\beta}(\boldsymbol{\theta})|\boldsymbol{\theta})) \left( \hat{g}_n^{(i)}(\mathbf{X}_i^\top \boldsymbol{\beta}(\boldsymbol{\theta})|\boldsymbol{\theta}) - g_0(\mathbf{X}_i^\top \boldsymbol{\beta}(\boldsymbol{\theta})|\boldsymbol{\theta}) \right)^m \\ & \quad + \bar{o} \left( n^\zeta \left( h_n^2 + \sqrt{\frac{\log n}{nh_n}} \right) \right) \\ &= \bar{o}(1). \end{aligned}$$

A similar argument shows that

$$\frac{1}{3!n} \sum_{i=1}^n \rho^{(3)}(Y_i, u_{ni}(\boldsymbol{\theta})) \left( \hat{g}_n^{(i)}(\mathbf{X}_i^\top \boldsymbol{\beta}(\boldsymbol{\theta})|\boldsymbol{\theta}) - g_0(\mathbf{X}_i^\top \boldsymbol{\beta}(\boldsymbol{\theta})|\boldsymbol{\theta}) \right)^3 \hat{\tau}_{ni}(\boldsymbol{\theta}) = \bar{o}(1).$$

As such, the first-order asymptotic behaviour of  $\sqrt{n}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0)$  is summarized by the representation

$$\sqrt{n}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0) = \frac{1}{\sqrt{n}} \sum_{i=1}^n \boldsymbol{\Psi}(\mathbf{Z}_i, \boldsymbol{\theta}_0, \hat{g}_n^{(i)}) + \bar{o}(1), \quad (31)$$

where for each  $i \in \{1, \dots, n\}$ ,

$$\begin{aligned} & \boldsymbol{\Psi}(\mathbf{Z}_i, \boldsymbol{\theta}_0, \hat{g}_n^{(i)}) \\ & \equiv \left[ \frac{1}{n} \sum_{j=1}^n \nabla_2 \rho(Y_j, \hat{g}_n^{(i)}(\mathbf{X}_j^\top \boldsymbol{\beta}(\boldsymbol{\theta}_0)|\boldsymbol{\theta}_0)) \hat{\tau}_{nj}(\boldsymbol{\theta}_0) \right]^{-1} \\ & \quad \cdot \nabla_1 \rho(Y_i, \hat{g}_n^{(i)}(\mathbf{X}_i^\top \boldsymbol{\beta}(\boldsymbol{\theta}_0)|\boldsymbol{\theta}_0)) \hat{\tau}_{ni}(\boldsymbol{\theta}_0). \end{aligned} \quad (32)$$

### A.3 Proof of Theorem 2

Start from the first-order representation given above in (31). Define for a smooth link function  $g : \mathbb{R} \rightarrow \mathbb{R}$  satisfying the conditions of Assumption 3 the functional

$$\psi(\mathbf{Z}_i, \boldsymbol{\theta}, g) \equiv \left[ \frac{1}{n} \sum_j \nabla_2 \rho(Y_j, g(\mathbf{X}_j^\top \boldsymbol{\beta}(\boldsymbol{\theta}) | \boldsymbol{\theta})) \right]^{-1} \nabla_1 \rho(Y_i, g(\mathbf{X}_i^\top \boldsymbol{\beta}(\boldsymbol{\theta}) | \boldsymbol{\theta})). \quad (33)$$

Let  $\psi_g$  and  $\psi_{gg}$  denote the first and second-order derivatives of  $\psi$  with respect to  $g$ . Expanding  $\psi(\mathbf{Z}_i, \boldsymbol{\theta}_0, \hat{g}_n^{(i)})$  about  $\psi(\mathbf{Z}_i, \boldsymbol{\theta}_0, g_0)$  to second order yields

$$\begin{aligned} & \sqrt{n}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0) \\ = & \frac{1}{\sqrt{n}} \sum_i \psi(\mathbf{Z}_i, \boldsymbol{\theta}_0, g_0) \\ & + \frac{1}{\sqrt{n}} \sum_i \psi_g(\mathbf{Z}_i, \boldsymbol{\theta}_0, g_0) \left( \hat{g}_n^{(i)}(\mathbf{X}_i^\top \boldsymbol{\beta}(\boldsymbol{\theta}_0) | \boldsymbol{\theta}_0) - G_0(\mathbf{X}_i^\top \boldsymbol{\beta}(\boldsymbol{\theta}_0)) \right) \\ & + \frac{1}{2\sqrt{n}} \sum_i \psi_{gg}(\mathbf{Z}_i, \boldsymbol{\theta}_0, g_0) \left( \hat{g}_n^{(i)}(\mathbf{X}_i^\top \boldsymbol{\beta}(\boldsymbol{\theta}_0) | \boldsymbol{\theta}_0) - G_0(\mathbf{X}_i^\top \boldsymbol{\beta}(\boldsymbol{\theta}_0)) \right)^2 \\ & + \bar{O}\left(n^\zeta \left(h_n^2 + \sqrt{\frac{\log n}{nh_n}}\right)\right) \\ = & \frac{1}{\sqrt{n}} \sum_i \psi(\mathbf{Z}_i, \boldsymbol{\theta}_0, g_0) \\ & + \frac{1}{\sqrt{n}} \sum_i \bar{\mu}_{gi}(\boldsymbol{\theta}_0, g_0) \left( \hat{g}_n^{(i)}(\mathbf{X}_i^\top \boldsymbol{\beta}(\boldsymbol{\theta}_0) | \boldsymbol{\theta}_0) - G_0(\mathbf{X}_i^\top \boldsymbol{\beta}(\boldsymbol{\theta}_0)) \right) \\ & + \frac{1}{\sqrt{n}} \sum_i \bar{\psi}_g(\mathbf{Z}_i, \boldsymbol{\theta}_0, g_0) \left( \hat{g}_n^{(i)}(\mathbf{X}_i^\top \boldsymbol{\beta}(\boldsymbol{\theta}_0) | \boldsymbol{\theta}_0) - G_0(\mathbf{X}_i^\top \boldsymbol{\beta}(\boldsymbol{\theta}_0)) \right) \\ & + \frac{1}{2\sqrt{n}} \sum_i \bar{\mu}_{ggi}(\boldsymbol{\theta}_0, g_0) \left( \hat{g}_n^{(i)}(\mathbf{X}_i^\top \boldsymbol{\beta}(\boldsymbol{\theta}_0) | \boldsymbol{\theta}_0) - G_0(\mathbf{X}_i^\top \boldsymbol{\beta}(\boldsymbol{\theta}_0)) \right)^2 \\ & + \frac{1}{2\sqrt{n}} \sum_i \bar{\psi}_{gg}(\mathbf{Z}_i, \boldsymbol{\theta}_0, g_0) \left( \hat{g}_n^{(i)}(\mathbf{X}_i^\top \boldsymbol{\beta}(\boldsymbol{\theta}_0) | \boldsymbol{\theta}_0) - G_0(\mathbf{X}_i^\top \boldsymbol{\beta}(\boldsymbol{\theta}_0)) \right)^2 \\ & + \bar{O}\left(n^\zeta \left(h_n^2 + \sqrt{\frac{\log n}{nh_n}}\right)\right), \end{aligned} \quad (34)$$

where

$$\begin{aligned} \bar{\mu}_{gi}(\boldsymbol{\theta}_0, g_0) & \equiv E \left[ \psi_g(\mathbf{Z}_i, \boldsymbol{\theta}_0, g_0) | \mathbf{X}_i \in \tilde{\Lambda}_n \right] \\ \bar{\psi}_g(\mathbf{Z}_i, \boldsymbol{\theta}_0, g_0) & \equiv \psi_g(\mathbf{Z}_i, \boldsymbol{\theta}_0, g_0) - \bar{\mu}_{gi}(\boldsymbol{\theta}_0, g_0) \\ \bar{\mu}_{ggi}(\boldsymbol{\theta}_0, g_0) & \equiv E \left[ \psi_{gg}(\mathbf{Z}_i, \boldsymbol{\theta}_0, g_0) | \mathbf{X}_i \in \tilde{\Lambda}_n \right] \\ \bar{\psi}_{gg}(\mathbf{Z}_i, \boldsymbol{\theta}_0, g_0) & \equiv \psi_{gg}(\mathbf{Z}_i, \boldsymbol{\theta}_0, g_0) - \bar{\mu}_{ggi}(\boldsymbol{\theta}_0, g_0). \end{aligned}$$

Consider

$$\mathbf{W}_1 \equiv \frac{1}{\sqrt{n}} \sum_i \bar{\mu}_{gi}(\boldsymbol{\theta}_0, g_0) \left( \hat{g}_n^{(i)}(\mathbf{X}_i^\top \boldsymbol{\beta}(\boldsymbol{\theta}_0) | \boldsymbol{\theta}_0) - G_0(\mathbf{X}_i^\top \boldsymbol{\beta}(\boldsymbol{\theta}_0)) \right).$$

Appealing to Lemma 4 we can write

$$\begin{aligned}
& \mathbf{W}_1 \\
&= \frac{1}{\sqrt{n}} \sum_i \bar{\mu}_{gi}(\boldsymbol{\theta}_0, g_0) \left[ \sum_{j \neq i} \epsilon_j(\boldsymbol{\theta}_0) w_{ij}(\boldsymbol{\theta}_0) + b_n(\mathbf{Z}_i | \boldsymbol{\theta}_0) \right] \\
&= \frac{1}{\sqrt{n}} \sum_i \bar{\mu}_{gi}(\boldsymbol{\theta}_0, g_0) \epsilon_i(\boldsymbol{\theta}_0) + \frac{1}{\sqrt{n}} \sum_i \bar{\mu}_{gi}(\boldsymbol{\theta}_0, g_0) \left[ \sum_{j \neq i} \epsilon_j(\boldsymbol{\theta}_0) w_{ij}(\boldsymbol{\theta}_0) - \epsilon_i(\boldsymbol{\theta}_0) \right] \\
&\quad + \frac{1}{\sqrt{n}} \sum_i \bar{\mu}_{gi}(\boldsymbol{\theta}_0, g_0) b_n(\mathbf{Z}_i | \boldsymbol{\theta}_0) \\
&\equiv \mathbf{W}_{11} + \mathbf{W}_{12} + \mathbf{W}_{13}.
\end{aligned} \tag{35}$$

Note that  $\mathbf{W}_{11} = \bar{O}(1)$  and is jointly asymptotically normal with the leading term

$$\mathbf{W}_0 \equiv \frac{1}{\sqrt{n}} \sum_i \psi(\mathbf{Z}_i, \boldsymbol{\theta}_0, g_0) \tag{36}$$

in (34).  $\mathbf{W}_{12}$  is mean zero and has variance proportional to

$$E \left[ \left( \sum_{j \neq i} \epsilon_j(\boldsymbol{\theta}_0) w_{ij}(\boldsymbol{\theta}_0) - \epsilon_i(\boldsymbol{\theta}_0) \right)^2 \right] = O(h_n^4).$$

Applying Lemma 1 it follows that  $\mathbf{W}_{12} = \bar{O}(h_n^2 \sqrt{\log n})$ . Finally note via Lemma 4 that  $\mathbf{W}_{13}$  has magnitude  $\bar{O}(\sqrt{n} h_n^2 + \sqrt{\frac{\log n}{n h_n}})$ . A further application of Lemma 4 yields the convergence

$$\begin{aligned}
& h_n^{-2} \cdot \frac{1}{n} \sum_i \bar{\mu}_{gi}(\boldsymbol{\theta}_0, g_0) b_n(\mathbf{Z}_i | \boldsymbol{\theta}_0) \\
& \xrightarrow{a.s.} E[\bar{\mu}_{g1}(\boldsymbol{\theta}_0, g_0) b_0(\mathbf{Z}_1 | \boldsymbol{\theta}_0)] \\
& \equiv \gamma_1,
\end{aligned} \tag{37}$$

where  $b_0(\mathbf{Z}_1 | \boldsymbol{\theta}_0)$  is as given above in the statement of Lemma 4.

Next, consider

$$\mathbf{W}_2 \equiv \frac{1}{\sqrt{n}} \sum_i \bar{\psi}_g(\mathbf{Z}_i, \boldsymbol{\theta}_0, g_0) \left( \hat{g}_n^{(i)}(\mathbf{X}_i^\top \boldsymbol{\beta}(\boldsymbol{\theta}_0) | \boldsymbol{\theta}_0) - G_0(\mathbf{X}_i^\top \boldsymbol{\beta}(\boldsymbol{\theta}_0)) \right).$$

We have

$$\begin{aligned}
& \mathbf{W}_2 \\
&= \frac{1}{\sqrt{n}} \sum_i \bar{\psi}_g(\mathbf{Z}_i, \boldsymbol{\theta}_0, g_0) \sum_{j \neq i} \epsilon_j(\boldsymbol{\theta}_0) w_{ij}(\boldsymbol{\theta}_0) + \frac{1}{\sqrt{n}} \sum_i \bar{\psi}_g(\mathbf{Z}_i, \boldsymbol{\theta}_0, g_0) b_n(\mathbf{Z}_i | \boldsymbol{\theta}_0) \\
&\equiv \mathbf{W}_{21} + \mathbf{W}_{22}.
\end{aligned}$$

Note that  $\mathbf{W}_{21}$  is mean zero, is uncorrelated with the leading terms and has variance of order  $O\left(\frac{1}{nh_n}\right)$ . Applying Lemma 1 we have  $\mathbf{W}_{21} = \bar{O}\left(\sqrt{\frac{\log n}{nh_n}}\right)$ .  $\mathbf{W}_{22}$  is also mean zero with variance of order  $O\left(\frac{1}{nh_n}\right)$ . A further application of Lemma 1 gives  $\mathbf{W}_{22} = \bar{O}\left(\sqrt{\frac{\log n}{nh_n}}\right)$ .

Now consider

$$\mathbf{W}_3 \equiv \frac{1}{2\sqrt{n}} \sum_i \bar{\mu}_{ggi}(\boldsymbol{\theta}_0, g_0) \left( \hat{g}_n^{(i)}(\mathbf{X}_i^\top \boldsymbol{\beta}(\boldsymbol{\theta}_0) | \boldsymbol{\theta}_0) - G_0(\mathbf{X}_i^\top \boldsymbol{\beta}(\boldsymbol{\theta}_0)) \right)^2.$$

We have via appeals to Lemma 1 and 4 that

$$\begin{aligned} & \mathbf{W}_3 \\ &= \frac{1}{2\sqrt{n}} \sum_i \bar{\mu}_{ggi}(\boldsymbol{\theta}_0, g_0) \left[ \sum_{j \neq i} \epsilon_j(\boldsymbol{\theta}_0) w_{ij}(\boldsymbol{\theta}_0) + b_n(\mathbf{Z}_i | \boldsymbol{\theta}_0) \right]^2 \\ &= \frac{1}{2\sqrt{n}} \sum_i \bar{\mu}_{ggi}(\boldsymbol{\theta}_0, g_0) \left[ \sum_{j \neq i} \epsilon_j(\boldsymbol{\theta}_0) w_{ij}(\boldsymbol{\theta}_0) \right]^2 + \bar{O}\left(\sqrt{nh_n^4} + \frac{\sqrt{\log n}}{nh_n}\right) \\ &= \frac{1}{2\sqrt{n}} \sum_i \bar{\mu}_{ggi}(\boldsymbol{\theta}_0, g_0) \sum_{j \neq i} E[\epsilon_j^2(\boldsymbol{\theta}_0) | \mathbf{X}_j] w_{ij}^2(\boldsymbol{\theta}_0) \\ &\quad + \frac{1}{2\sqrt{n}} \sum_i \bar{\mu}_{ggi}(\boldsymbol{\theta}_0, g_0) \sum_{j \neq i} w_{ij}^2(\boldsymbol{\theta}_0) [\epsilon_j^2(\boldsymbol{\theta}_0) - E[\epsilon_j^2(\boldsymbol{\theta}_0) | \mathbf{X}_j]] \\ &\quad + \frac{1}{2\sqrt{n}} \sum_i \bar{\mu}_{ggi}(\boldsymbol{\theta}_0, g_0) \underbrace{\sum_{j \neq 1} \sum_{k \neq 1} w_{ij}(\boldsymbol{\theta}_0) w_{ik}(\boldsymbol{\theta}_0) \epsilon_j(\boldsymbol{\theta}_0) \epsilon_k(\boldsymbol{\theta}_0)}_{j \neq i, k} + \bar{O}\left(\sqrt{nh_n^4} + \frac{\sqrt{\log n}}{nh_n}\right) \\ &\equiv \mathbf{W}_{31} + \mathbf{W}_{32} + \mathbf{W}_{33} + \bar{O}\left(\sqrt{nh_n^4} + \frac{\sqrt{\log n}}{nh_n}\right). \end{aligned}$$

For

$$\sigma^2(\boldsymbol{\theta}_0) \equiv E[\epsilon_1^2(\boldsymbol{\theta}_0)]$$

we have via an appeal to Lemma 1 that

$$\mathbf{W}_{31} = \frac{\sigma^2(\boldsymbol{\theta}_0) \int K^2(s) ds}{2h_n \sqrt{n}} \cdot \frac{1}{n} \sum_i \bar{\mu}_{ggi}(\boldsymbol{\theta}_0, g_0) \cdot \frac{1}{f_0(\mathbf{X}_i^\top \boldsymbol{\beta}(\boldsymbol{\theta}_0) | \boldsymbol{\theta}_0)} + \bar{o}\left(\frac{\sqrt{\log n}}{h_n \sqrt{n}}\right).$$

As such,

$$\mathbf{W}_{31} = \frac{1}{h_n \sqrt{n}} \gamma_2 + \bar{o}\left(\frac{\sqrt{\log n}}{h_n \sqrt{n}}\right),$$

where

$$\gamma_2 \equiv \sigma^2(\boldsymbol{\theta}_0) \int K^2(s) ds \cdot E\left[\bar{\mu}_{gg1}(\boldsymbol{\theta}_0, g_0) \cdot \frac{1}{f_0(\mathbf{X}_1^\top \boldsymbol{\beta}(\boldsymbol{\theta}_0) | \boldsymbol{\theta}_0)}\right]. \quad (38)$$

Noting that both  $\mathbf{W}_{32}$  and  $\mathbf{W}_{33}$  are mean zero, similar appeals to Lemma 1 yield  $\mathbf{W}_{32} = \bar{O}\left(\frac{\sqrt{n}}{nh_n}\right)$  and  $\mathbf{W}_{33} = \bar{O}\left(\sqrt{\frac{\log n}{nh_n}}\right)$ .

Finally consider

$$\mathbf{W}_4 \equiv \frac{1}{2\sqrt{n}} \sum_i \bar{\psi}_{gg}(\mathbf{Z}_i, \boldsymbol{\theta}_0, g_0) \left( \hat{g}_n^{(i)}(\mathbf{X}_i^\top \boldsymbol{\beta}(\boldsymbol{\theta}_0) | \boldsymbol{\theta}_0) - G_0(\mathbf{X}_i^\top \boldsymbol{\beta}(\boldsymbol{\theta}_0)) \right)^2.$$

Note that  $\bar{\psi}_{gg}(\mathbf{Z}_i, \boldsymbol{\theta}_0, g_0)$  is conditionally independent of

$$\left( \hat{g}_n^{(i)}(\mathbf{X}_i^\top \boldsymbol{\beta}(\boldsymbol{\theta}_0) | \boldsymbol{\theta}_0) - G_0(\mathbf{X}_i^\top \boldsymbol{\beta}(\boldsymbol{\theta}_0)) \right)^2$$

given  $\mathbf{X}_i$ . As such,  $\mathbf{W}_4$  is mean zero, and also uncorrelated with the asymptotically normal terms in (34), namely  $\mathbf{W}_0$  and  $\mathbf{W}_{11}$  as given above in (36) and (35), respectively. Deduce via Lemmas 4 and 1 that

$$\begin{aligned} \mathbf{W}_4 &= \frac{1}{2\sqrt{n}} \sum_i \bar{\psi}_{gg}(\mathbf{Z}_i, \boldsymbol{\theta}_0, g_0) \left( \sum_{j \neq i} \epsilon_j(\boldsymbol{\theta}_0) w_{ij}(\boldsymbol{\theta}_0) + b_n(\mathbf{Z}_i | \boldsymbol{\theta}_0) \right)^2 \\ &= \bar{O}\left(\sqrt{nh_n^4} + \frac{\sqrt{\log n}}{nh_n}\right). \end{aligned}$$

Summarizing, we have the representation

$$\sqrt{n}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0) = \mathbf{A}_{n0} + \mathbf{A}_{n1} + \mathbf{A}_{n2} + \mathbf{A}_{n3} + \mathbf{A}_{n4} + \mathbf{R}_n.$$

Here

$$\begin{aligned} \mathbf{A}_{n0} &\equiv \frac{1}{\sqrt{n}} \sum_i (\psi(\mathbf{Z}_i, \boldsymbol{\theta}_0, g_0) + \bar{\mu}_{gi}(\boldsymbol{\theta}_0, g_0) \epsilon_i(\boldsymbol{\theta}_0)) \\ &= \bar{O}(1) \end{aligned} \tag{39}$$

and is asymptotically  $N(\mathbf{0}, \boldsymbol{\Sigma}_0)$ , where

$$\boldsymbol{\Sigma}_0 \equiv E \left[ (\psi(\mathbf{Z}_1, \boldsymbol{\theta}_0, g_0) + \bar{\mu}_{g1}(\boldsymbol{\theta}_0, g_0) \epsilon_1(\boldsymbol{\theta}_0)) (\psi(\mathbf{Z}_1, \boldsymbol{\theta}_0, g_0) + \bar{\mu}_{g1}(\boldsymbol{\theta}_0, g_0) \epsilon_1(\boldsymbol{\theta}_0))^\top \right]; \tag{40}$$

while

$$\begin{aligned} &\mathbf{A}_{n1} \\ &\equiv \frac{1}{\sqrt{n}} \sum_i \bar{\mu}_{gi}(\boldsymbol{\theta}_0, g_0) b_n(\mathbf{Z}_i | \boldsymbol{\theta}_0) \\ &= \sqrt{nh_n^2} \boldsymbol{\gamma}_1 + \bar{o}(1). \end{aligned} \tag{41}$$

In addition,

$$\begin{aligned} &\mathbf{A}_{n2} \\ &\equiv \frac{\sigma^2(\boldsymbol{\theta}_0) \int K^2(s) ds}{2h_n \sqrt{n}} \cdot \frac{1}{n} \sum_i \bar{\mu}_{ggi}(\boldsymbol{\theta}_0, g_0) \cdot \frac{1}{f_0(\mathbf{X}_i^\top \boldsymbol{\beta}(\boldsymbol{\theta}_0) | \boldsymbol{\theta}_0)} \\ &= \frac{1}{h_n \sqrt{n}} \boldsymbol{\gamma}_2 + \bar{o}(1), \end{aligned} \tag{42}$$

$$\begin{aligned}
& \equiv \frac{\mathbf{A}_{n3}}{\sqrt{n}} \sum_i \left[ \bar{\psi}_g(\mathbf{Z}_i, \boldsymbol{\theta}_0, g_0) \left( \hat{g}_n^{(i)}(\mathbf{X}_i^\top \boldsymbol{\beta}(\boldsymbol{\theta}_0) | \boldsymbol{\theta}_0) - G_0(\mathbf{X}_i^\top \boldsymbol{\beta}(\boldsymbol{\theta}_0)) \right) \right. \\
& \quad \left. + \frac{1}{2} \bar{\mu}_{ggi}(\boldsymbol{\theta}_0, g_0) \sum_{\substack{j \neq 1 \\ j \neq i, k}} \sum_{k \neq 1} w_{ij}(\boldsymbol{\theta}_0) w_{ik}(\boldsymbol{\theta}_0) \epsilon_j(\boldsymbol{\theta}_0) \epsilon_k(\boldsymbol{\theta}_0) \right] \\
& = \bar{O} \left( \sqrt{\frac{\log n}{nh_n}} \right),
\end{aligned} \tag{43}$$

$$\begin{aligned}
& \equiv \frac{\mathbf{A}_{n4}}{\sqrt{n}} \sum_i \left\{ \bar{\mu}_{gi}(\boldsymbol{\theta}_0, g_0) \left[ \sum_{j \neq i} \epsilon_j(\boldsymbol{\theta}_0) w_{ij}(\boldsymbol{\theta}_0) - \epsilon_i(\boldsymbol{\theta}_0) \right] \right. \\
& \quad + \frac{1}{2} \bar{\mu}_{ggi}(\boldsymbol{\theta}_0, g_0) \sum_{j \neq i} w_{ij}^2(\boldsymbol{\theta}_0) [\epsilon_j^2(\boldsymbol{\theta}_0) - E[\epsilon_j^2(\boldsymbol{\theta}_0) | \mathbf{X}_j]] \\
& \quad \left. + \frac{1}{2} \bar{\psi}_{gg}(\mathbf{Z}_i, \boldsymbol{\theta}_0, g_0) \left[ \hat{g}_n^{(i)}(\mathbf{X}_i^\top \boldsymbol{\beta}(\boldsymbol{\theta}_0) | \boldsymbol{\theta}_0) - G_0(\mathbf{X}_i^\top \boldsymbol{\beta}(\boldsymbol{\theta}_0)) \right]^2 \right\} \\
& = \bar{O} \left( h_n^2 \sqrt{\log n} + \frac{\sqrt{\log n}}{nh_n} + \sqrt{nh_n^4} \right)
\end{aligned} \tag{44}$$

and

$$\mathbf{R}_n = \bar{O} \left( n^\zeta \left( h_n^2 + \sqrt{\frac{\log n}{nh_n}} \right) \right).$$

For bandwidth  $h_n \propto n^{-\frac{1}{3}}$  and trimming parameter  $\zeta \in (0, \frac{2}{15})$  we have

$$\begin{aligned}
\mathbf{A}_{n1} &= n^{-\frac{1}{6}} \boldsymbol{\gamma}_1 + \bar{o}(1), \\
\mathbf{A}_{n2} &= n^{-\frac{1}{6}} \boldsymbol{\gamma}_2 + \bar{o}(1), \\
\mathbf{A}_{n3} &= \bar{O} \left( n^{-\frac{1}{3}} \sqrt{\log n} \right), \\
\mathbf{A}_{n4} &= \bar{O} \left( n^{-\frac{2}{3}} \sqrt{\log n} \right)
\end{aligned}$$

and

$$\mathbf{R}_n = \bar{o} \left( n^{-\frac{1}{5}} \sqrt{\log n} \right).$$

It follows that under the conditions of Assumptions 1–6,  $\sqrt{n}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0)$  has the representation

$$\sqrt{n}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0) = \mathbf{A}_{n0} + \sqrt{n} h_n^2 \boldsymbol{\gamma}_1 + \frac{1}{h_n \sqrt{n}} \boldsymbol{\gamma}_2 + \bar{o} \left( n^{-\frac{1}{5}} \sqrt{\log n} \right). \tag{45}$$

## A.4 Proof of Theorem 3

Recall the notation defined above the statement of Lemma 5 above that  $\mathbf{A}_m^* = \bar{O}^*(a_m)$  (or  $\bar{o}^*(a_m)$ ) iff each element of  $\mathbf{A}_m^*$  is uniformly  $O_{a.s.}(a_m)$  (resp.  $o_{a.s.}(a_m)$ ) over  $\Lambda_m$  and  $\Theta_n$  conditionally on  $\mathcal{X}_n$  as  $m, n \rightarrow \infty$ . For

$$\epsilon_{i,m}^*(\boldsymbol{\theta}) = Y_{i,m}^* - \hat{g}_n^{(i,m)}(\mathbf{X}_{i,m}^{*\top} \boldsymbol{\beta}(\boldsymbol{\theta}) | \boldsymbol{\theta}), \quad (46)$$

let

$$\sigma^{*2}(\boldsymbol{\theta}) \equiv Var[\epsilon_{i,m}^*(\boldsymbol{\theta}) | \mathcal{X}_n] \quad (47)$$

and let  $\bar{\sum}_{(i,m)}$  denote summation over observations in  $\mathcal{X}_m^*$  with  $\mathbf{X}_{i,m}^* \in \tilde{\Lambda}_m$ .

An application of Lemma 5 and imitation of the steps taken in Appendix A.3 yields the following representation for  $\sqrt{m}(\hat{\boldsymbol{\theta}}_m^* - \hat{\boldsymbol{\theta}}_n)$  when  $h_m \propto m^{-\frac{1}{3}}$  and  $\zeta \in (0, \frac{2}{15})$ :

$$\begin{aligned} \sqrt{m}(\hat{\boldsymbol{\theta}}_m^* - \hat{\boldsymbol{\theta}}_n) &= \frac{1}{\sqrt{m}} \bar{\sum}_{(i,m)} \left( \psi^*(\mathbf{Z}_{i,m}^*, \hat{\boldsymbol{\theta}}_n, \hat{g}_n) + \bar{\boldsymbol{\mu}}_{gi}^*(\hat{\boldsymbol{\theta}}_n, \hat{g}_n) \epsilon_{i,m}^*(\hat{\boldsymbol{\theta}}_n) \right) \\ &\quad + \frac{1}{\sqrt{m}} \bar{\sum}_{(i,m)} \bar{\boldsymbol{\mu}}_{gi}^*(\hat{\boldsymbol{\theta}}_n, \hat{g}_n) b_m^*(\mathbf{Z}_{i,m}^* | \hat{\boldsymbol{\theta}}_n) \\ &\quad + \frac{\sigma^{*2}(\hat{\boldsymbol{\theta}}_n) \int K^2(s) ds}{2h_m \sqrt{m}} \cdot \frac{1}{\sqrt{m}} \bar{\sum}_{(i,m)} \bar{\boldsymbol{\mu}}_{ggi}^*(\hat{\boldsymbol{\theta}}_n, \hat{g}_n) \cdot \frac{1}{\hat{f}_n^{(i,m)}(\mathbf{X}_{i,m}^{*\top} \boldsymbol{\beta}(\hat{\boldsymbol{\theta}}_n) | \hat{\boldsymbol{\theta}}_n)} \\ &\quad + \bar{o}\left(m^{-\frac{1}{5}} \sqrt{\log m}\right), \end{aligned} \quad (48)$$

where

$$\begin{aligned} \psi^*(\mathbf{Z}_{i,m}^*, \hat{\boldsymbol{\theta}}_n, \hat{g}_n) &\equiv \left[ \frac{1}{m} \bar{\sum}_{(j,m)} \nabla_2 \rho(Y_{j,m}^*, \hat{g}_n^{(i,m)}(\mathbf{X}_{j,m}^{*\top} \boldsymbol{\beta}(\hat{\boldsymbol{\theta}}_n) | \hat{\boldsymbol{\theta}}_n)) \right]^{-1} \\ &\quad \cdot \nabla_1 \rho(Y_{i,m}^*, \hat{g}_n^{(i,m)}(\mathbf{X}_{i,m}^{*\top} \boldsymbol{\beta}(\hat{\boldsymbol{\theta}}_n) | \hat{\boldsymbol{\theta}}_n)); \\ \bar{\boldsymbol{\mu}}_{gi}^*(\hat{\boldsymbol{\theta}}_n, \hat{g}_n) &\equiv E[\psi_g(\mathbf{Z}_{i,m}^*, \hat{\boldsymbol{\theta}}_n, \hat{g}_n) | \mathbf{X}_{i,m}^* \in \tilde{\Lambda}_m, \mathcal{X}_n] \end{aligned}$$

and

$$\bar{\boldsymbol{\mu}}_{ggi}^*(\hat{\boldsymbol{\theta}}_n, \hat{g}_n) \equiv E[\psi_{gg}(\mathbf{Z}_{i,m}^*, \hat{\boldsymbol{\theta}}_n, \hat{g}_n) | \mathbf{X}_{i,m}^* \in \tilde{\Lambda}_m, \mathcal{X}_n].$$

Analyze the first three terms in (48) in succession.

First consider

$$\begin{aligned} \mathbf{W}_{m0}^* &\equiv \frac{1}{\sqrt{m}} \bar{\sum}_{(i,m)} \left( \psi^*(\mathbf{Z}_{i,m}^*, \hat{\boldsymbol{\theta}}_n, \hat{g}_n) + \bar{\boldsymbol{\mu}}_{gi}^*(\hat{\boldsymbol{\theta}}_n, \hat{g}_n) \epsilon_{i,m}^*(\hat{\boldsymbol{\theta}}_n) \right) \\ &\equiv \frac{1}{\sqrt{m}} \bar{\sum}_{(i,m)} \mathbf{W}_{m0i}^*. \end{aligned} \quad (49)$$

Note that  $\mathbf{W}_{m0}^* = \bar{O}^*(1)$  and is asymptotically  $N(E[\mathbf{W}_{m0i}^* | \mathcal{X}_n], Var[\mathbf{W}_{m0i}^* | \mathcal{X}_n])$  conditional on  $\mathcal{X}_n$ . But

$$\begin{aligned} E[\mathbf{W}_{m0i}^* | \mathcal{X}_n] &\xrightarrow{a.s.} E[\psi(\mathbf{Z}_1, \boldsymbol{\theta}_0, g_0) + \bar{\boldsymbol{\mu}}_{g1}(\boldsymbol{\theta}_0, g_0) \epsilon_1(\boldsymbol{\theta}_0)] \\ &= \mathbf{0} \end{aligned}$$



and

$$\text{Var}[\mathbf{W}_{m0i}^* | \mathcal{X}_n] \xrightarrow{a.s.} \Sigma_0,$$

where  $\Sigma_0$  is as given above in (40). It follows that  $\mathbf{W}_{m0}^* = \mathbf{A}_{m0} + \bar{o}(1)$ , where  $\mathbf{A}_{m0}$  is as given above in (39) with  $n = m$ .

Next consider

$$\begin{aligned} \mathbf{W}_{m1}^* &\equiv \frac{1}{\sqrt{m}} \sum_{(i,m)} \bar{\mu}_{gi}^* (\hat{\theta}_n, \hat{g}_n) b_m^* (\mathbf{Z}_{i,m}^* | \hat{\theta}_n) \\ &\equiv \frac{1}{\sqrt{m}} \sum_{(i,m)} \mathbf{W}_{m1i}^*. \end{aligned} \quad (50)$$

When  $m \in [C_1 n^\xi, C_2 n^{1-\xi}]$  for suitable positive constants  $C_1, C_2$  and  $\xi$ , an application of Lemma 5 yields

$$\begin{aligned} h_m^{-2} \cdot \frac{1}{m} \sum_{(i,m)} \mathbf{W}_{m1i}^* &= h_m^{-2} E \left[ \bar{\mu}_{gi}^* (\hat{\theta}_n, \hat{g}_n) b_m^* (\mathbf{Z}_{i,m}^* | \hat{\theta}_n) \middle| \mathcal{X}_n \right] + \bar{o}^*(1) \\ &\xrightarrow{a.s.} E [\bar{\mu}_{g1}(\theta_0, g_0) b_0(\mathbf{Z}_1 | \theta_0)] \\ &= \gamma_1, \end{aligned}$$

where  $\gamma_1$  is as given above in (37).

On the other hand, a setting of  $m = n$  yields via another application of Lemma 5 the representation

$$\mathbf{W}_{m1}^* = \bar{O} \left( h_n^2 + \sqrt{\frac{\log n}{n h_n}} \right) = \bar{o} \left( m^{-\frac{1}{5}} \sqrt{\log m} \right).$$

It follows that

$$\mathbf{W}_{m1}^* = h_m^2 \sqrt{m} \gamma_1 + \bar{o}(1)$$

iff  $m \in [C_1 n^\xi, C_2 n^{1-\xi}]$ ,  $n \rightarrow \infty$  and  $h_m \propto m^{-\frac{1}{3}}$ .

Finally, take

$$\begin{aligned} \mathbf{W}_{m2}^* &\equiv \frac{\sigma^{*2}(\hat{\theta}_n) \int K^2(s) ds}{2h_m \sqrt{m}} \cdot \frac{1}{m} \sum_{(i,m)} \bar{\mu}_{ggi}^* (\hat{\theta}_n, \hat{g}_n) \cdot \frac{1}{\hat{f}_n^{(i,m)} (\mathbf{X}_{i,m}^{*\top} \boldsymbol{\beta}(\hat{\theta}_n) | \hat{\theta}_n)} \\ &\equiv \frac{\sigma^{*2}(\hat{\theta}_n) \int K^2(s) ds}{2h_m \sqrt{m}} \cdot \frac{1}{m} \sum_{(i,m)} \mathbf{W}_{m2i}^*. \end{aligned} \quad (51)$$

We have as  $m, n \rightarrow \infty$

$$\begin{aligned} \frac{1}{m} \sum_{(i,m)} \mathbf{W}_{m2i}^* &= E \left[ \bar{\mu}_{ggi}^* (\hat{\theta}_n, \hat{g}_n) \cdot \frac{1}{\hat{f}_n^{(i,m)} (\mathbf{X}_{i,m}^{*\top} \boldsymbol{\beta}(\hat{\theta}_n) | \hat{\theta}_n)} \middle| \mathcal{X}_n \right] + \bar{o}^*(1) \\ &\xrightarrow{a.s.} E \left[ \bar{\mu}_{gg1}(\theta_0, g_0) \cdot \frac{1}{f_0(\mathbf{X}_1^\top \boldsymbol{\beta}(\theta_0) | \theta_0)} \right] \\ &= \frac{1}{\sigma^2(\theta_0) \int K^2(s) ds} \gamma_2, \end{aligned}$$

where  $\gamma_2$  is as given above in (38). It follows that

$$\mathbf{W}_{m2}^* = \frac{1}{h_m \sqrt{m}} \gamma_2 + \bar{o}(1)$$

as  $m, n \rightarrow \infty$ .

Summarizing, when  $h_m \in [\delta m^{-\frac{1}{3}}, \delta^{-1} m^{-\frac{1}{3}}]$  for some small  $\delta \in (0, 1)$  and  $m \in [C_1 n^\xi, C_2 n^{1-\xi}]$  and  $\zeta \in (0, \frac{2}{15})$  the representation

$$\sqrt{m} (\hat{\boldsymbol{\theta}}_m^* - \hat{\boldsymbol{\theta}}_n) = \mathbf{A}_{m0} + h_m^2 \sqrt{m} \gamma_1 + \frac{1}{h_m \sqrt{m}} \gamma_2 + \bar{o} \left( m^{-\frac{1}{5}} \sqrt{\log m} \right).$$

This shows that  $\sqrt{m} (\hat{\boldsymbol{\theta}}_m^* - \hat{\boldsymbol{\theta}}_n)$  can be made almost surely equal in large samples to  $\sqrt{m} (\hat{\boldsymbol{\theta}}_m - \boldsymbol{\theta}_0)$  up to terms of order  $m^{-\frac{1}{6}}$ .

On the other hand, a setting of  $m = n$  under the same constraints on  $h_m$  yields the representation

$$\sqrt{m} (\hat{\boldsymbol{\theta}}_m^* - \hat{\boldsymbol{\theta}}_n) = \mathbf{A}_{m0} + \frac{1}{h_m \sqrt{m}} \gamma_2 + \bar{o} \left( m^{-\frac{1}{5}} \sqrt{\log m} \right).$$

## A.5 Proof of Theorem 4

Let  $\mathbf{W}_{m0}^*$ ,  $\mathbf{W}_{m1}^*$  and  $\mathbf{W}_{m2}^*$  be as given above in (49), (50) and (51), respectively. Define

$$\mathbf{W}_m^* \equiv \mathbf{W}_{m0}^* + \mathbf{W}_{m1}^* + \mathbf{W}_{m2}^*.$$

It follows from Theorem 3 that choosing the bandwidth  $h_{m,opt}^*$  to minimize

$$E \left[ m (\hat{\boldsymbol{\theta}}_m^* - \hat{\boldsymbol{\theta}}_n)^\top \boldsymbol{\Sigma}_0^{-1} (\hat{\boldsymbol{\theta}}_m^* - \hat{\boldsymbol{\theta}}_n) \middle| \mathcal{X}_n \right]$$

is asymptotically equivalent to defining  $h_{m,opt}^*$  as the bandwidth value that minimizes  $E [\mathbf{W}_m^* | \mathcal{X}_n]$ . By Lemmas 1 and 2 we have that

$$E [\mathbf{W}_{m0}^* | \mathcal{X}_n] = \bar{O} \left( \sqrt{m} \left( \sqrt{\frac{\log n}{n}} + h_n^2 + \sqrt{\frac{\log n}{n h_n}} \right) \right).$$

On the other hand,  $E [\mathbf{W}_{m1}^* | \mathcal{X}_n] = \bar{O} (h_m^2 \sqrt{m})$  and  $E [\mathbf{W}_{m2}^* | \mathcal{X}_n] = \bar{O} \left( \frac{1}{h_m \sqrt{m}} \right)$ . From this it follows that the magnitude of  $E [\mathbf{W}_m^* | \mathcal{X}_n] - E [\mathbf{W}_m]$ , where

$$\mathbf{W}_m \equiv \mathbf{A}_{m0} + h_m^2 \sqrt{m} \gamma_1 + \frac{1}{h_m \sqrt{m}} \gamma_2,$$

is of order  $\sqrt{\frac{m \log n}{n}} + m^{-\frac{1}{6}}$  when  $h_m \propto m^{-\frac{1}{3}}$  and  $E [\mathbf{W}_{m1}^* | \mathcal{X}_n]$  and  $E [\mathbf{W}_{m2}^* | \mathcal{X}_n]$  are both  $\bar{O} (m^{-\frac{1}{6}})$ .

It follows that setting  $m \propto \left( \frac{n}{\log n} \right)^{\frac{3}{4}}$  is sufficient to minimize the order of the estimation error involved in the substitution of  $E [\mathbf{W}_m^* | \mathcal{X}_n]$  for  $E [\mathbf{W}_m]$ .

Figure 1: Relative root MSE as a function of bandwidth

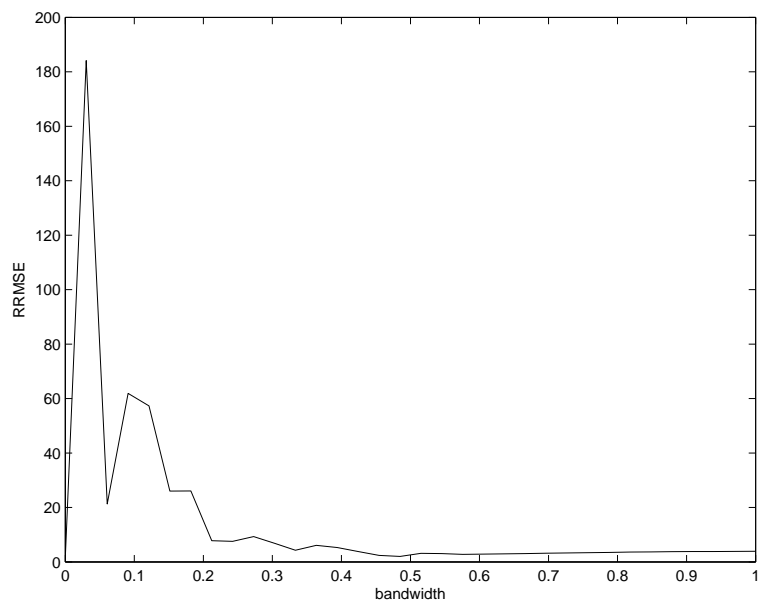


Figure 2: Relative root MSE as a function of bandwidth, log-log scale

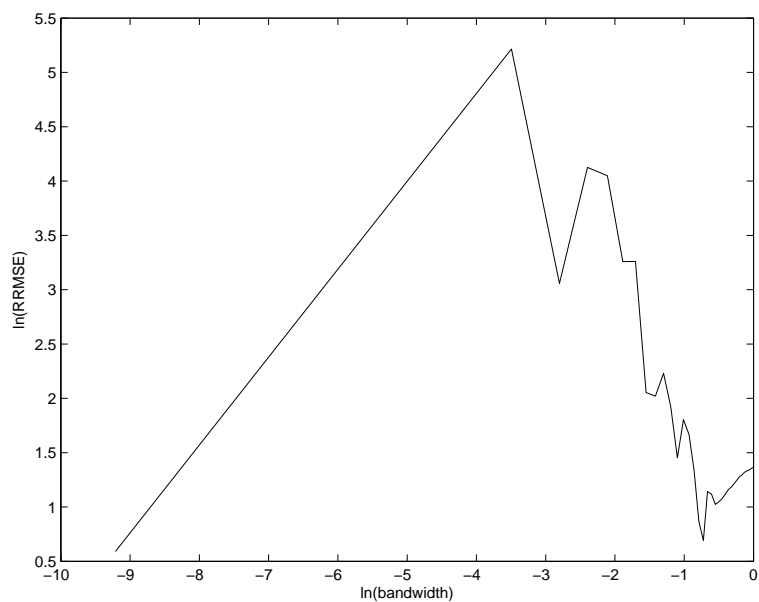


Table 1: Finite-sample behaviour of the bandwidth estimators,  $n = 1000$

Bandwidth	Mean	Standard deviation	RMSE
$h_{n,opt}$	.4785	—	—
$\hat{h}_{\lfloor m^* \rfloor, opt}$	.6220	.3502	.3781
$\hat{h}_{\lfloor \frac{3}{2} m^* \rfloor, opt}$	.3350	.1858	.2346
$\hat{h}_{\lfloor 2m^* \rfloor, opt}$	.2243	.1070	.2758
$\hat{h}_{\lfloor 5m^* \rfloor, opt}$	.1914	.1592	.3282
$\hat{h}_{n, Sil}$	.9743	.0627	.4997

Table 2: Finite-sample behaviour of the SLS estimator with various bandwidths,  $n = 1000$

Estimator	Mean	SD	RMSE
$\hat{\theta}_n(h_{n,opt})$	-.7283	.0897	.2861
$\hat{\theta}_n(\hat{h}_{\lfloor m^* \rfloor, opt})$	-.7125	.0949	.3027
$\hat{\theta}_n(\hat{h}_{\lfloor \frac{3}{2} m^* \rfloor, opt})$	-.7286	.0941	.2872
$\hat{\theta}_n(\hat{h}_{\lfloor 2m^* \rfloor, opt})$	-.7269	.0899	.2875
$\hat{\theta}_n(\hat{h}_{\lfloor 5m^* \rfloor, opt})$	-.7143	.0811	.2970
$\hat{\theta}_n(\hat{h}_{n, Sil})$	-.6938	.1249	.3306